Accepted Manuscript

Performance ranking (dis)similarities in commodity markets

Hanxiong Zhang, Benjamin R. Auer, Dimitrios I. Vortelinos

PII: S1044-0283(17)30143-6
DOI: doi: 10.1016/j.gfj.2017.09.001
Reference: GLOFIN 384

To appear in:

Received date: 2 May 2017
Revised date: 5 September 2017
Accepted date: 11 September 2017

Performance ranking (dis)similarities in commodity markets

Hanxiong Zhang,[a] Benjamin R. Auer,[b] Dimitrios I. Vortelinos[c]

[a]Lincoln International Business School, University of Lincoln
Brayford Pool
Lincoln
LN6 7TS
United Kingdom
Telephone: 0044-1522-835639
hzhang@lincoln.ac.uk

[b]Corresponding author
University of Leipzig, Department of Finance
Grimmaische Straße 12
04109 Leipzig
Germany
Telephone: 0049-341-9733672
auer@wifa.uni-leipzig.de

[c]Lincoln International Business School, University of Lincoln
Brayford Pool
Lincoln
LN6 7TS
United Kingdom
Telephone: 0044-1522-835634
dvortelinos@lincoln.ac.uk

ABSTRACT
In this article, we revisit recent evidence indicating that the choice of performance measure appears to be irrelevant for the ranking of investment alternatives in the commodity market. Extending the previous literature in several important ways, we provide the following insights into the rankings produced by the 13 most popular performance measures for 24 commodities. First, ranking differences are somewhat larger in the spot market than in the futures market. Second, when we use daily instead of monthly data, performance measures that model reward based on average returns still produce similar performance rankings. However, when data of higher frequency is used for performance measures modeling reward based on higher partial moments, performance rankings differ crucially from those produced by measures focusing on average returns. Finally, the degree of ranking (dis)similarity appears to vary over time. Empirically, then, the choice of performance measure can matter. Nevertheless, our findings do not invalidate recent theoretical results on ranking similarity, because population rankings may not be identical with sample rankings, which are subject to estimation error.

JEL classifications:
C10
D81
G11
G29

Keywords:
Performance ranking
Commodity investments
Data frequency
Market phase dependency

1. Introduction

With their influential studies, Eling and Schuhmacher (2007) and Eling (2008) started an ongoing debate on whether the choice of performance measure matters in the evaluation of asset performance.[1] For a wide variety of investment fund datasets, they document high rank correlations between the Sharpe ratio and several alternative reward-to-risk ratios based on drawdowns, partial moments, and the Value-at-Risk. Because this result suggests that investors could prefer simpler performance measures to more complex ones, the findings of these studies are of high practical importance and have quickly stimulated further research.

Zakamouline (2011) reinvestigates the findings of Eling and Schuhmacher (2007) for hedge funds by taking a more detailed look at the rankings produced by different performance measures instead of focusing just on rank correlations. By calculating the maximum upgrade, maximum downgrade, mean absolute change, and standard deviation of the change in the rankings, he argues that a high rank correlation coefficient does not necessarily imply almost identical rank orders, because there are funds that show substantial changes in ranking if the performance measure is changed from the Sharpe ratio to an alternative measure.[2] Ornelas, Silva Júnior, and Fernandes (2012) reinvestigate the findings of Eling (2008) for mutual funds and suggest that performance measures do not yield similar rankings if their reward measures are different (e.g., when the mean excess return is replaced

[1] Cogneau and Hübner (2015) summarize the results of earlier studies which have received relatively little attention because of their small fund samples (e.g., Gemmill, Hwang, & Salmon, 2005; Hwang & Salmon, 2002; Plantinga & De Groot, 2001) and their limited selection of performance measures (e.g., Frohlich, Schnusenberg, & Pennathur, 2006).

[2] Adcock, Areal, Armada, Cortez, Oliveira and Silva (2015) report similar results for a sample of UK investment trusts.

by a higher partial moment).[3] Eling, Farinelli, Rossello, and Tibiletti (2011), looking at the Sharpe ratio and several performance measures based on partial moments (the Sortino-Satchell, Farinelli-Tibiletti, and Rachev ratios), argue that the choice of performance measure in hedge fund evaluations is irrelevant only when these alternative measures are tailored to a moderate investment style. When they are used to describe aggressive investment styles, rank correlations with the Sharpe ratio decrease significantly. Finally, Auer and Schuhmacher (2013) use a selection of more advanced rank correlation measures and find that adequately defined drawdown-based performance measures yield hedge fund rankings that are not too different from those of the Sharpe ratio when investors are primarily interested in picking the best investments and when a sufficiently large return sample is used to calculate performance measure estimates. They also highlight that the rankings are not strictly identical when small return samples are analyzed.

While most studies in this field have concentrated on rankings of investment funds, the recent contribution of Auer (2015a) focuses on commodity markets, where the Sharpe ratio has become the dominant measure for evaluating and comparing different commodity trading strategies (Bianchi, Drew, & Fan, 2015a, 2015b; Erb & Harvey, 2006; Fuertes, Miffre, & Fernandez-Perez, 2015; Fuertes, Miffre, & Rallis, 2010; Gorton & Rouwenhorst, 2006; Miffre & Rallis, 2007; Szakmary, Shen, & Sharma, 2010) and the question of whether one type of investment is superior to another may be answered differently when other performance measures are employed.[4] Using a sample of 24 highly liquid commodity

[3] They also show that the rankings of the less frequently used “manipulation-proof performance measure” and the “appraisal ratio” differ considerably from Sharpe ratio rankings.

[4] Pedersen and Rudholm-Alfvin (2003) and Caporin and Lisi (2011) also do not focus on funds; they concentrate on stocks. The former study reports that, for symmetric return

futures, Auer (2015a) shows that the Sharpe ratio and its 12 most popular alternatives yield almost identical rankings of investment alternatives. He also shows that his empirical findings are robust to changes in the futures dataset, the use of equal-length subsamples, and the performance measure parameterization.

Given that the findings of Eling and Schuhmacher (2007) and Eling (2008) for investment funds have been challenged by several follow-up studies, our goal is to analyze whether the commodity market results of Auer (2015a) have general validity. This kind of analysis is important because Jensen (1967) argues that reinvestigation based on additional bodies of data for other time-periods is among the best ways to refute charges of data mining. Specifically, we provide three important contributions to the literature. First, in contrast to the majority of studies in the field, we use daily data instead of monthly data.[5] Second, we analyze a potential dependency of ranking similarities on the market phase by dividing our sample into subsamples classified by prevalent market conditions. Finally, in addition to the futures market data of Auer (2015a), we also look at the (partially hypothetical) performance of commodity spot market investments.

The remainder of this article is organized as follows. In Section 2, we derive our research hypotheses from empirical and theoretical results of earlier studies. Section 3 reports the main features of our dataset. Section 4 briefly describes the selection of performance measures (and their specifications) used in our study. Section 5 reports our

---

distributions, rank correlation between different performance measures is high, whereas there is a significant absence of consistency when the distribution of returns is asymmetric. It also shows that correlation is higher for financial firms and lower for small firms. The latter study indicates that rank correlations may vary over time.

[5] Ornelas et al. (2012) and Adcock et al. (2015) are the only researchers who have used daily data (of investment funds) to analyze performance ranking similarities; other authors use monthly data only.

empirical results and verbally summarizes the outcomes of several robustness checks. Section 6 concludes by discussing our empirical results in the context of recent theoretical literature on ranking similarities and by pointing out directions for future research with global relevance.

2. Research hypotheses

For each of our contributions stated in Section 1, we can formulate a hypothesis reflecting the result we might expect in our empirical commodity market analysis.

Starting with our focus on data of higher frequency, it is well known that the accuracy of risk measure estimates improves greatly with the sample frequency (Burghardt & Walls, 2011; Frazzini & Pedersen, 2014) and that a smaller number of observations means that there will be less or even no information available about the extreme tails of a return distribution (Adcock et al., 2015). Furthermore, as different kinds of risk measures have distinct degrees of estimation error in small samples (Schuhmacher & Auer, 2014) and thus react differently to changes in the sample frequency, a change from monthly to daily data may have crucial impact on empirically observable similarities in commodity rankings. This leads us to our first research hypothesis.

Hypothesis 1. Using data of higher frequency leads to larger ranking differences.

As far as our subsample analysis is concerned, previous research has shown that the moments of asset returns (Jondeau & Rockinger, 2003) and correlations between asset returns (Ang & Bekaert, 2002) are not only time-varying but strongly driven by the general market direction (Amira, Taamouti, & Tsafack, 2011). Furthermore, Krimm, Scholz, and

Wilkens (2012) highlight a significant impact of market climates on Sharpe ratios.[6] Thus, unless different performance measures are affected by market climate in exactly the same way, we might expect that rank correlations between return-based performance measures also vary over time, so that we may observe alternating periods of stronger and weaker similarities in rankings. This expectation forms our second research hypothesis.

Hypothesis 2. The degree of ranking (dis)similarity varies over time.

Finally, we look at the commodity dataset used in the performance evaluation. Auer (2015a) focuses on commodity futures returns because investors can easily realize such returns either by trading the futures contracts at almost negligible transaction costs or by investing in exchange traded commodities (ETCs) tracking the futures. We are interested in whether switching to another kind of commodity investment crucially influences the similarity among performance rankings. Specifically, in addition to futures data, we have a look at the commodity spot market. Here, one could realize spot market returns by physical investment. However, in practice, investors do not choose such a course of action for most commodities (except for precious metals) because they are typically unable to store the purchased quantities (e.g., natural gas). Instead, investors hoping to directly capture spot market movements invest in certificates and ETCs, which are collateralized by holdings in the physical commodity and thus deliver spot returns (minus costs of storage) to the investors.[7]

[6] They show that the Sharpe ratios of poorly diversified investment funds are biased upwards in bear markets, and vice versa. Subsequently, they confirm that fund Sharpe ratios depend on especially the mean excess return (over the risk-free rate) of the market.

[7] Such products are typically not available for perishable commodities (e.g., agricultural

Even though basic theory tells us that futures and spot markets should be strongly linked through the cost-of-carry model (Hull, 2006; Pindyck, 2001), empirical results often suggest otherwise. In the crude oil market, for example, several studies find only limited evidence of cointegration and Granger-causality between the two markets (Bekiros & Diks, 2008), and others report that futures market data have almost no power to predict spot market prices (Alquist & Kilian, 2010). It is not surprising, then, that the two markets tend to show quite different return characteristics—for example, for extreme observations (Arouri, Lahiani, Lévy, & Nguyen, 2012). This allows us to hypothesize that the degree of ranking similarity may differ depending on the market.

Hypothesis 3. The level of ranking (dis)agreement differs in spot and futures markets.

In what follows, we analyze whether the empirical evidence supports or disconfirms these hypotheses.

3. Data

3.1. Data source and return calculation

This study employs data from the constituents (subindices) of the Standard and Poor's Goldman Sachs Commodity Index (S&P GSCI) for the period from January 7, 2002, to March 31, 2016.[8] This index covers 24 commodities from a wide variety of sectors: six

---

products and livestock). Thus, our investment rankings based on spot market data are partially hypothetical. Furthermore, they abstract from the costs of storage. Nonetheless, even such hypothetical rankings provide important information for comparing general market developments in different commodity segments.

[8] Even though some of the subindices have a longer data history, we select a consistent start

energy products (Brent crude oil, WTI crude oil, gas oil, heating oil, natural gas, unleaded gasoline), two precious metals (gold, silver), five industrial metals (aluminum, copper, lead, nickel, zinc), eight agricultural products (cocoa, coffee, corn, cotton, soybeans, sugar, Chicago wheat, Kansas wheat), and three livestock products (feeder cattle, lean hogs, live cattle). For each commodity, we obtain a total return futures index and a spot index. The futures index measures the returns accrued from investing in liquid fully collateralized nearby futures, whereas the spot index reflects the performance of a physical commodity investment. The main advantage of the futures index is that it is completely comparable to returns from a regular investment in the S&P 500 (with dividend reinvestment) or a government bond.[9] This is why these indices have become popular benchmarks for evaluating investment strategies (e.g., momentum trading rules) in commodity futures markets (Bianchi et al., 2015a; Erb & Harvey, 2006) and important tools for identifying the best investment opportunity in a set of given commodity alternatives (Auer, 2015a).

The daily data for the futures and spot indices are obtained from Thomson Reuters Datastream. Following the standard convention (Miffre & Rallis, 2007), we calculate daily log returns as $r_t = (ln\ I_t - ln\ I_{t-1}) \cdot 100$, where I_t is the index value for a given commodity on day t and ln denotes a natural logarithm. Excess returns are computed by subtracting the daily U.S. Treasury bill rate of Ibbotson Associates (archived by Kenneth French at mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html).

in January 2002 because adequately comparing investment performance across all indices requires using the same sample period with the same number of observations for each index.

[9] For details on the index construction (e.g., the procedure used to roll over from one futures contract to the next), see www.goldmansachs.com/what-we-do/securities/products-and-business-groups/products/gcsi.

3.2. Characteristics of the full sample

Table 1 presents the minimum, maximum, mean, standard deviation, skewness, kurtosis, and results of the Jarque and Bera (1987) test for normality, for the commodity excess returns in our sample period.

[Insert Table 1 about here]

In the futures market, the highest daily losses and gains can be observed for the energy, precious metals, and industrial metals sectors (in line with findings by Auer, 2014; Doran & Ronn, 2008; Sévi, 2015). For example, silver and nickel (natural gas and WTI crude oil) exhibit the most significant losses (gains), of -19.49% and -18.26% (18.77% and 13.34%), respectively. In contrast, the livestock sector is more stable with respect to such drastic outliers, partly because it has been less subject to speculative attacks than the others (Auer, 2015a). Turning to the mean excess returns, we find the highest (lowest) value for copper (natural gas), i.e., 0.04% (-0.14%). Standard deviations take their highest (lowest) values, 2.99% (0.92%), for natural gas (feeder cattle). With the exceptions of gas oil, natural gas, coffee, corn, and wheat, all excess return distributions are negatively skewed, suggesting that there is a higher chance of realizing high negative excess returns than large positive ones. All commodities have kurtosis values larger than 3, indicating heavier tails and/or stronger peaks than those in a normal distribution. Given these properties, the null hypothesis of normally distributed excess returns is rejected for all commodities.[10] A brief look at the results for the spot returns yields largely similar results.

[10] These strong deviations from normality are usually the reason why many researchers discard the Sharpe ratio and resort to alternative performance measures (Eling & Schuhmacher, 2007).

3.3. Subsample properties

In order to analyze whether ranking similarities are linked to (i) the data frequency used for performance evaluation and/or (ii) the market phase in which the performance is measured, we divide our sample into seven subsamples.[11]

Auer subsample: January 7, 2002–September 30, 2013. This sample period has been used by Auer (2015a). He used monthly data and found that the choice of performance measure was largely irrelevant. Thus, this subsample will help determine whether this result also holds for data of higher frequency.

S1—the Argentina crisis subsample: January 7, 2002–November 30, 2002. In this period, international markets were affected by the Argentine crisis in December 2001, when the government of Argentina declared itself unable to pay its debts (Cho, Hyde, & Nguyen, 2015).

S2—the growth subsample: December 1, 2002–August 1, 2008. This was a period of economic growth with low inflation, significant international trade, and large financial flows in emerging and developing countries (see http://www.nber.org/cycles).

S3—the Lehman Brothers crisis subsample: September 1, 2008–December 7, 2010. This period covers the expansion of the Federal Reserve Bank (FED) and European Central Bank (ECB) balance sheets because of liquidity problems that seized financial markets following the collapse of Lehman Brothers (Cukierman, 2013; Shachmurove, 2011).

S4—the EU crisis subsample: December 8, 2010–April 4, 2011. This period starts with the beginning of the EU debt crisis and ends with its most consequential period (Cho et al., 2015).

[11] Some of our subsamples cover economic crises and downturns in stock or bond markets (Xing, 2017). Using more commodity-related subsamples (as did Adams & Glück, 2015; Belousova & Dorfleitner, 2012; Bianchi, Drew, & Fan, 2016) leads to similar results.

S5—the Greek crisis subsample: April 5, 2011–March 31, 2012. This period covers the peak of the Greek sovereign crisis, in which the ECB's rate of balance sheet expansion accelerated by 70.88% per annum (Cukierman, 2013).

S6—the post-crisis subsample: April 1, 2012–March 31, 2016. This period is characterized by a relaxation of the Greek sovereign crisis and lasts until the end of our sample.

Figure 1 (2) summarizes the descriptive statistics for the commodity futures (spot) excess returns in our subsamples; in each subfigure, we concentrate on one descriptive measure (mean, standard deviation, skewness, or kurtosis). To understand our specific form of visualization, take the standard deviation as an example. In the corresponding subfigure, we plot the mean of the 24 commodity standard deviations for each subsample and also report their minimum and maximum as a band around the mean value. This way we can visualize the evolution of commodity return standard deviations over our subsamples and also illustrate dispersion across the commodities.

[Insert Figures 1 and 2 about here]

A look at the futures returns in Figure 1 shows that the characteristics of the Auer subsample do not differ crucially from those of our full sample. Among our subsamples S1 to S6, the EU crisis subsample S4 was the best period for commodity investments in terms of average mean excess returns. It also had relatively low average commodity market volatility. In contrast, the Lehman subsample S3 shows rather low (and on average negative) excess returns and significantly higher average volatility. Skewness and kurtosis appear in all of our subsamples and vary over time, as do higher moments of other asset classes (Jondeau & Rockinger, 2003; León, Rubio, & Serna, 2005).

Apart from somewhat lower dispersion in the means, the descriptive statistics of the spot subsamples in Figure 2 reveal results largely similar to those shown in Figure 1.

4. Methods

In our study, we focus on a selection of performance measures most popular in research and practice (Eling & Schuhmacher, 2007; Auer, 2015a). These 13 reward-to-risk measures (defined in Table 2) mainly share the same reward measure (the mean excess return) in the numerator but differ with respect to the type of risk measure in the denominator. They fall into four main groups: “classic”, “based on drawdowns”, “based on partial moments”, and “based on the Value-at-Risk.”

[Insert Table 2 about here]

The Sharpe ratio is one of the most popular performance measures in the investment industry (Eling & Schuhmacher, 2007; Shukla & Singh, 1997). For a long time, researchers mistakenly believed that the measure had a decision theoretic foundation only in cases where returns were normally distributed (Auer, 2015c). This erroneous belief and the measure’s technical defects (i.e., its vulnerability to option-based manipulation strategies or distortions introduced by very high or very low returns; see Auer, 2013; Goetzmann, Ingersoll, Spiegel, & Welch, 2007; Schuster & Auer, 2012) made researchers and practitioners look for alternatives.

Given the drawback that the standard deviation used as a risk measure in the Sharpe ratio treats positive deviations from the mean as risks, straightforward modifications of the Sharpe ratio employ risk measures focusing on worst-case events. In this spirit, a first class of alternative performance measures uses drawdowns to quantify risk. We use five measures of this class, namely, the Calmar ratio, Sterling ratio, Burke ratio, Pain ratio, and Martin ratio (as defined by Schuhmacher & Eling, 2011). While the Calmar ratio quantifies risk as the maximum drawdown, the Sterling and Burke ratios use the mean and the square root of the sum of squares (which puts a stronger emphasis on large losses) of the *K* largest continuous drawdowns, respectively. We follow the literature standard and set *K=5* (Auer & Schuhmacher, 2013; Auer, 2015a). Finally, the Pain ratio and Martin ratio quantify risk by calculating the mean of the percentage drops from the previous peak and the square root of the mean of the squared percentage drops from the previous peak, respectively. This procedure takes into account the duration of drawdowns.

Another class of alternative performance measures uses partial moments to quantify risk (and reward). In contrast to the standard deviation, lower partial moments focus only on negative deviations from a minimal acceptable excess return (which is zero in our definition, following Schuhmacher & Eling, 2012). Again, we select the four most popular metrics of this kind, namely, the Omega ratio, Sortino ratio, Kappa 3 ratio, and upside potential ratio. The former three measures use the (normalized) lower partial moments of orders one, two, and three, respectively, where a higher order models more risk-averse investors (Eling & Schuhmacher, 2007; Eling et al., 2011). While these measures use the mean excess return to quantify reward, the upside potential ratio modifies the Sortino ratio by replacing the mean excess return with the higher partial moment of order one (focusing on positive deviations from the minimal acceptable return).

Finally, three ratios are based on the Value-at-Risk (VaR). The VaR used in the excess return on VaR quantifies the possible percentage loss of an investment, which is not exceeded with a given probability $1-\alpha$ in a certain period. The conditional VaR used in the conditional Sharpe ratio represents the expected percentage loss if the VaR is exceeded. We estimate both risk measures using historical simulation because it can account for non-normally distributed returns (Auer, 2015b) and because it is the most popular method in the industry (Pérignon & Smith, 2010).[12] Finally, the modified Sharpe ratio makes use of the

[12] In a survey of the VaR disclosures of 60 U.S., Canadian, and large international banks from 1996 to 2005, Pérignon and Smith (2010) document that 73% of the banks used the

modified VaR, an extension of the standard VaR formula for normal distributions that accounts for skewness and kurtosis in the data (Eling, 2008). For all of our VaR-based measures, we set $\alpha = 5\%$, which is a typical value in this context (Gilli & Këllezi, 2006).

Given that negative mean excess returns can have a distorting influence on asset rankings, we apply the Israelsen (2005) correction. That is, for each performance measure with reward measure θ and risk measure φ, we do not use θ/φ when ranking commodities but use the measure $\theta/\varphi^{\theta/abs(\theta)}$. For positive excess returns, this formula is identical to the original performance measure θ/φ. If the excess return is negative, we get the expression $\theta\varphi$, yielding a correct ranking.

5. Empirical analysis

5.1. Commodity rankings and rank correlations

In our empirical investigation, we take the perspective of a commodity investor who has access to the dataset presented in Section 3 and is interested in identifying the best commodity investments by evaluating historical performance.

For our full sample, Table 3 presents the rankings generated by each of our performance measures as well as a mean ranking across all measures. Subdivided into futures- and spot-based investments, the commodities are ranked from best (rank 1) to worst (rank 24).

[Insert Table 3 about here]

We observe the following. First, while rankings in futures markets are widely similar across performance measures that measure reward by means of average returns, the rankings derived from the upside potential ratio, which measures reward by means of higher partial moments, are crucially different. For example, while gold is the best investment according to

historical simulation method.

the average return measures (and according to Caporin, Ranaldo, & Velo, 2015; Kristjanpoller & Minutolo, 2015; O'Connor, Lucey, Batten, & Baur, 2015), it reaches only rank 10 according to the upside potential ratio. Second, in spot markets, the upside potential ratio is similarly exceptional. Also, the differences in rankings produced by the other performance measures are larger in spot data than in futures data. Finally, the mean ranks (across all performance measures) in futures and spot markets are quite different. For example, while gold, soybeans, and copper (natural gas, lean hogs, and Chicago wheat) are the three best (worst) investments according to futures data, gold, copper, and coffee (aluminum, natural gas, and nickel) are the best (worst) investments according to spot data. This indicates that, as hypothesis 3 predicts, the choice of database (futures vs. spot) can crucially affect performance evaluation and ranking differences between alternative performance measures.

A similar picture emerges in the results for our subsamples (see Tables A1 to A7 of the appendix). In addition, they provide the following insights. First, when comparing our results based on daily data in the Auer subsample to the original results of Auer (2015a) based on monthly data, we find that using a different data frequency influences the ranking outcome, as hypothesis 1 predicts.[13] For example, while the upside potential ratio produces rankings similar to the other measures with monthly data, it does not do so with daily data. Second, our results show that the relative performance of commodities varies over time.[14] For example, while in futures data the Sharpe ratio ranks gold fifth in our growth subsample

[13] This result is in line with Christoffersen and Langlois (2013) showing that the choice of data frequency can crucially influence results in the context of factor model estimation.

[14] This is reasonable because previous studies have shown that asset risk tends to vary over time (see, for example, Bollerslev, 1986; Chatziantoniou, Filis, & Floros, 2017; Degiannakis & Floros, 2016; Engle, 1982; Ohlson & Rosenberg, 1982; Pan, Liu, & Roth, 1999).

(roughly, 2002–2008), the same ratio ranks it twentieth in the EU crisis subsample (roughly, the first quarter of 2011). A closer inspection of the subsample results also reveals that the general market direction (boom or bust) is insufficient to explain these variations. This may be partly because commodity prices are no longer determined simply by demand and supply but now also by the continuing financialization of the commodity market (Tang & Xiong, 2012). Finally and most importantly, not only the relative performance of commodities but also the difference in the rankings produced by different performance measures appears to vary over time. Thus, we cannot reject hypothesis 2. We focus on this point in the remainder of our analysis by introducing some compact measures of ranking similarity which can be easily compared across subsamples.

We start by computing Kendall's τ and Spearman's ρ rank correlation coefficients, which are typical measures for such a purpose (Auer & Schuhmacher, 2013). The main difference between the two measures is that, in the calculation of Spearman's ρ, large differences in rankings have higher weights than small differences, whereas Kendalls's τ does not consider the severity of differences but concentrates on whether or not there are differences at all. Large values of both measures indicate strong ranking similarities, and a value of one reflects equality of the two rankings used to calculate the rank correlation coefficients.

Because the Sharpe ratio is the simplest of our performance measures, it is typically used as the benchmark. This means that rank correlations are calculated between this measure and potential alternatives (Ornelas et al., 2012; Zakamouline, 2011). Table 4 reports these rank correlations for our full sample; it focuses on the results based on futures data because, currently, ETCs based on futures are available in a wider variety than ETCs capturing spot market prices. Thus our futures-based results have higher practical relevance

(Garner, 2012).[15]

[Insert Table 4 about here]

Table 4 suggests that the ranks delivered by the Sharpe ratio are highly correlated with those generated by all alternative measures using mean returns to measure reward, a pattern that is consistent with the findings of Auer and Schuhmacher (2013) and Auer (2015a). Correlations vary from 0.9904 (Pain ratio) to 0.9983 (Omega ratio, Sortino ratio, Kappa 3 ratio, excess return on VaR) and from 0.9348 (Pain ratio) to 0.9855 (Omega ratio, Sortino ratio, Kappa 3 ratio, excess return on VaR) according to Spearman's ρ and Kendall's τ, respectively. However, the rank correlations between the Sharpe ratio and the upside potential ratio are significantly lower. Here, we have $\rho = 0.2552$ and $\tau = 0.1884$, values that are in line with the results of Zakamouline (2011) and Ornelas et al. (2012) for monthly data. That is, investors using the upside potential ratio obtain investment decisions that differ crucially from decisions obtained using other performance measures.

Figure 3, which follows the design of Figures 1 and 2, allows judging the evolution of rank correlations over time. For each of our subsamples, we plot the minimum, maximum, and average of all rank correlations with the Sharpe ratio. Because of the distinct role of the upside potential ratio, we create separate plots including and excluding the correlation values for this performance measure.

[Insert Figure 3 about here]

The mean of the rank correlations is crucially affected by the results for the upside potential ratio. After it is excluded, all mean rank correlations (regardless of the type of rank correlation coefficient) are close to one. Furthermore, time-varying rank correlations suggest that the degree of similarity in ranking varies over time. While the rank correlations for the measures using mean return reward vary only a little over time, the correlations for the

[15] Detailed results for the spot market data are available from the authors upon request.

upside potential ratio change more significantly between periods.

To address the criticism of Zakamouline (2011) and Ornelas et al. (2012) that high rank correlations do not necessarily imply almost identical rankings, Table 4 presents the descriptive statistics for the differences in ranks (minimum differences, maximum differences, mean absolute differences, and standard deviation of absolute differences) between the 12 alternative performance measures and the Sharpe ratio. Supporting our rank correlation analysis, rankings do not drastically change when an alternative performance measure using mean return reward is applied instead of the Sharpe ratio. In the most extreme case, the Calmar ratio, one commodity moves down 2 (minimum of -2) places and another one moves up 2 (maximum of 2) places. However, when we look at the upside potential ratio, on average a commodity moves 6.42 places. This high difference is also reflected by a high standard deviation of absolute differences, taking a value of 5.61.

Figure 4, which summarizes the mean absolute differences and the standard deviation of mean absolute differences for our subperiods, paints a similar picture. As in Figure 3, we find that the upside potential ratio contributes significantly to the mean of ranking differences across performance measures and that the magnitude of its deviations from the Sharpe ratio differs depending on the subsample. Interestingly, all performance measures show their strongest deviations from the Sharpe ratio in the most recent subsample.

[Insert Figure 4 about here]

5.2. Focus on best-performing commodities

Because investors are typically interested in identifying the best investments, several studies have suggested that a focus on these investments may produce additional insights into ranking similarities (Auer & Schuhmacher, 2013; Zakamouline, 2011).[16] Following this

[16]Of course, if investors would like to implement momentum strategies which require short-

suggestion, Table 5 presents the ranking difference statistics for the five commodities with the highest Sharpe ratio in the full sample. That is, we identify the commodities with the highest Sharpe ratios, rank these commodities from best (rank 1) to worst (rank 5) according to the Sharpe ratio and the alternative performance measures, and then use these ranks to calculate the ranking differences.[17]

[Insert Table 5 about here]

Interestingly, a focus on the best investments drastically reduces the differences in rankings for all alternative performance measures, including the upside potential ratio. In the most extreme cases, ranks are now changed by only one position when switching to another performance measure. In the majority of cases, the ranks are not changed at all, so that mean absolute differences and standard deviations of absolute differences fall below 1. A similar picture emerges when we repeat this analysis for our subsamples. Figure 5 shows that even though ranking differences do not seriously influence decision making, in some periods (e.g., the EU crisis, subsample S4) the rankings of the alternative measures differ more from the Sharpe ratio rankings than they do in others.

[Insert Figure 5 about here]

---

selling the commodities with the poorest past performance (see, e.g., Miffre & Ralllis, 2007; Szakmary et al., 2010), the worst investments also become relevant. Therefore, we extended our analysis to the ranking differences among the worst five commodities. The results are similar to those for the top five commodities, suggesting that ranking similarities are stronger for “extreme performers” than for “average performers”.

[17] In contrast, Auer and Schuhmacher (2013) calculate the differences using the original ranks of the alternative performance measures in the ranking of all investment alternatives.

5.3. Some final robustness checks

While the robustness of our results with respect to different datasets and subsamples has already been part of our main analysis, this section covers some additional aspects of parameter choice in the calculation of the performance measures. We follow Eling and Schuhmacher (2007) and Eling (2008) by varying the significance level α in the VaR-based measures between 1% and 10% in steps of one and the number of drawdowns K in the Sterling and Burke ratios between 1 and 10 in steps of one. However, these changes do not influence our overall results on ranking (dis)similarity between these measures and the Sharpe ratio.

6. Conclusion

In this article, we reexamine Auer's (2015a) finding that the choice of performance measure does not crucially affect the rankings of investment alternatives in commodity markets. Specifically, we analyze whether his result holds for (i) data of higher frequency, (ii) subsamples reflecting different market phases, and (iii) alternative data (spot market returns). Our results show that switching from monthly to daily data widens the ranking differences among our 13 performance measures. Especially, the upside potential ratio, which measures investment reward using higher partial moments, generates crucially different rankings than do performance metrics that use mean excess returns. This holds in both futures and spot market data, with larger differences in the latter case. Furthermore, ranking differences appear to vary over time—again, especially for the upside potential ratio. Finally, supporting earlier literature, we also find that ranking disagreement is considerably smaller for the top five commodities than for all alternatives.

Even though these results challenge previous studies arguing that in typical empirical applications the choice of performance measure is irrelevant, our findings are still in line with recent studies developing the theoretical conditions under which different performance measures produce identical rankings of risky alternatives. Schuhmacher and Eling (2011, 2012) show that if investment returns fulfill the location and scale (LS) condition of Sinn (1983) and Meyer (1987), identical rankings are produced by the Sharpe ratio, by adequately defined drawdown-based performance measures, and by certain performance measures based on partial moments, the VaR, and other risk quantities. Given that the LS condition cannot be satisfied in an environment with cross-sectionally different levels of skewness and kurtosis, Schuhmacher and Auer (2014) show that these performance measures also yield identical rankings when the generalized LS condition of Meyer and Rasche (1992) holds, which allows for cross-sectional differences in skewness and kurtosis.

To relate our results to this literature, we have to consider that, in empirical studies, we are working with small samples, while the theoretical literature refers to population properties. Thus, even if the generalized LS condition holds and we have identical population rankings, the rankings in small samples may still be different because of estimation error, which differs for each performance measure (Schuhmacher & Auer, 2014). Thus, our detected ranking differences do not challenge the theoretical literature on ranking similarities. To challenge the empirical relevance this literature, we would have to show that the generalized LS condition does not hold or at least has weaker empirical support in specific subsamples. This would deliver a perfect explanation for the time variation in ranking similarities that we have detected. Unfortunately, the statistical techniques currently available do not allow testing the generalized LS condition, so that more work on adequate statistical methods is required to answer this question (Auer, 2015c).

While this aspect of our work offers plenty of scope for future theoretical research, our findings also suggest directions for additional empirical research. First, recent work analyzing global commodity market investments has strongly focused on the performance of futures-based momentum, reversal, and term structure strategies (Bianchi et al., 2015a, 2015b; Fuertes et al. 2010, 2015). Here, the typical approach is to evaluate different strategies in different subsamples by looking at their Sharpe ratios. But our results indicate that evaluation can yield different outcomes when another performance measure is used instead of the Sharpe ratio. Thus, from a practical perspective, it may be interesting to expand our study of individual commodities to ranking advanced commodity trading strategies (involving more than one commodity) and their different specifications. Second, a similar reinvestigation might be performed in another context. Recent studies have shown that commodity trading advisors, which manage a crucial proportion of global commodity investments, display poor performance against simple benchmarks (Bhardwaj, Gorton, & Rouwenhorst, 2014). This raises the question of whether such a judgement also holds under the entire set of alternative performance measures. Third, as far as global portfolio management is concerned, experimental evidence suggests that even skilled investors often ignore correlations in asset allocation decisions (Kallir & Sonsino, 2009) and base their actions mainly on stand-alone information for individual assets (as we do in our analysis). Consequently, it would be interesting to see how the weights of commodities in international (diversified) portfolios change under alternative measures of individual performance. Changes in investment weights caused by persuasively promoting alternatives to the Sharpe ratio may have non-negligible impacts on markets, since significant capital flows of financial investors in commodity futures markets have been shown to affect the dynamics of futures markets (Fattouh, Kilian, & Mahadeva, 2013).

Because the empirical finding that various performance measures generate similar performance rankings appears to be not very robust, investors have two options. First, they may simply use more than one performance measure. This way they can evaluate different aspects of the performance of their trading strategies. If investment alternatives must be ranked, investors could use the performance measure that is consistent with their individual utility function (Zakamouline, 2011, 2014). Second, investors may extract the common information contained in all available performance measures and use this information for ranking purposes. This can be achieved by applying principal-component analysis within a large universe of performance measure specifications (Cogneau & Hübner, 2015).

Acknowledgements

We thank an anonymous reviewer for valuable comments and suggestions. We are also indebted to the Fritz Thyssen Stiftung (grant 20.15.0.079 WW) for generous financial support.

References

Adams, Z., & Glück, T. (2015). Financialization in commodity markets: A passing trend or the new normal? *Journal of Banking and Finance*, *60*, 93–111.

Adcock, C. J., Areal, N., Armada, M. J. R., Cortez, M. C., Oliveira, B., & Silva, F. (2015). *Portfolio performance measurement: Monotonicity with respect to the Sharpe ratio and multivariate tests of correlation*. Unpublished manuscript, University of Sheffield.

Alquist, R., & Kilian, L. (2010). What do we learn from the price of crude oil futures? *Journal of Applied Econometrics*, *25*(4), 539–573.

Amira, K., Taamouti, A., & Tsafack, G. (2011). What drives international equity correlations? Volatility or market direction? *Journal of International Money and Finance*, *30*(6), 1234–1263.

Ang, A., & Bekaert, G. (2002). International asset allocation with regime shifts. *Review of Financial Studies*, *15*(4), 1137–1187.

Arouri, M. E. H., Lahiani, A., Lévy, A., & Nguyen, D. K. (2012). Forecasting the conditional volatility of oil spot and futures prices with structural breaks and long memory models. *Energy Economics*, *34*(1), 283–293.

Auer, B. R. (2013). The low return distortion of the Sharpe ratio. *Financial Markets and Portfolio Management*, *27*(3), 299–306.

Auer, B. R. (2014). Could diamonds become an investor's best friend? *Review of Managerial Science*, *8*(3), 351–383.

Auer, B. R. (2015a). Does the choice of performance measure influence the evaluation of commodity investments? *International Review of Financial Analysis*, *38*, 142–150.

Auer, B. R. (2015b). Extreme value theory, asset ranking and threshold choice: A practical note on VaR estimation. *Journal of Risk*, *18*(1), 27–44.

Auer, B. R. (2015c). On the role of skewness, kurtosis, and the location and scale condition in a Sharpe ratio performance evaluation setting. *International Journal of Theoretical and Applied Finance*, *18*(6), 1550037:1–13.

Auer, B. R., & Schuhmacher, F. (2013). Robust evidence on the similarity of Sharpe ratio and drawdown-based hedge fund performance rankings. *Journal of International Financial Markets, Institutions and Money*, *24*, 153–165.

Bekiros, S. D., & Diks, C. G. H. (2008). The relationship between crude oil spot and futures prices: Cointegration, linear and nonlinear causality. *Energy Economics*, *30*(5), 2673–2685.

Belousova, J., & Dorfleitner, G. (2012). On the diversification benefits of commodities from the perspective of Euro investors. *Journal of Banking and Finance*, *36*(9), 2455–2472.

Bhardwaj, G., Gorton, G. B., & Rouwenhorst, K. G. (2014). Fooling some of the people all of the time: The inefficient performance and persistence of commodity trading advisors. *Review of Financial Studies*, *27*(11), 3099–3132.

Bianchi, R. J., Drew, M. E., & Fan, J. H. (2015a). Combining momentum with reversal in commodity futures. *Journal of Banking and Finance*, *59*, 423–444.

Bianchi, R. J., Drew, M. E., & Fan, J. H. (2015b). *Microscopic momentum in commodity futures*. Griffith University Working Paper No. 2015-10.

Bianchi, R. J., Drew, M. E., & Fan, J. H. (2016). Commodities momentum: A behavioural perspective. *Journal of Banking and Finance*, *72*, 133–150.

Bollerslev, T. (1986). Generalised autoregressive conditional heteroskedasticity. *Journal of Econometrics*, *31*(3), 307–327.

Burghardt, G., & Walls, B. (2011). *Managed futures for institutional investors: Analysis and portfolio construction*. Hoboken: Wiley.

Caporin, M., & Lisi, F. (2011). Comparing and selecting performance measures using rank correlations. *Economics: The Open-Access, Open-Assessment E-Journal*, *5*(2011-10), 1–

34.

Caporin, M., Ranaldo, A., & Velo, G. (2015). Precious metals under the microscope: A high-frequency analysis. *Quantitative Finance*, *15*(5), 743–759.

Chatziantoniou, I., Filis, G., & Floros, C. (2017). Asset prices regime-switching and the role of inflation targeting monetary policy. *Global Finance Journal*, *32*, 97–112.

Cho, S., Hyde, S., & Nguyen, N. (2015). Time-varying regional and global integration and contagion: Evidence from style portfolios. *International Review of Financial Analysis*, *42*, 109–131.

Christoffersen, P., & Langlois, H. (2013). The joint dynamics of equity market factors. *Journal of Financial and Quantitative Analysis*, *48*(5), 1371–1404.

Cogneau, P., & Hübner, G. (2015). *The dimensions of mutual funds performance and persistence*. Unpublished manuscript, Maastricht University.

Cukierman, A. (2013). Monetary policy and institutions before, during, and after the global financial crisis. *Journal of Financial Stability*, *9*(3), 373–384.

Degiannakis, S., & Floros, C. (2016). Intra-day realized volatility for European and USA stock indices. *Global Finance Journal*, *29*, 24–41.

Doran, J. S., & Ronn, E. I. (2008). Computing the market price of volatility risk in the energy commodity markets. *Journal of Banking and Finance*, *32*(12), 2541–2552.

Eling, M. (2008). Does the measure matter in the mutual fund industry? *Financial Analysts Journal*, *64*(3), 54–66.

Eling, M., Farinelli, S., Rossello, D., & Tibiletti, L. (2011). One-size or tailor-made performance ratios for ranking hedge funds? *Journal of Derivatives and Hedge Funds*, *16*(4), 267–277.

Eling, M., & Schuhmacher, F. (2007). Does the choice of performance measure influence the evaluation of hedge funds? *Journal of Banking and Finance*, *31*(9), 2632–2647.

Engle, R. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, *50*(4), 987–1007.

Erb, C. B., & Harvey, C. R. (2006). The strategic and tactical value of commodity futures. *Financial Analysts Journal*, *62*(2), 69–97.

Fattouh, B., Kilian, L., & Mahadeva, L. (2013). The role of speculation in oil markets: What have we learned so far? *Energy Journal*, *34*(3), 7–33.

Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, *111*(1), 1–25.

Frohlich, C., Schnusenberg, O., & Pennathur, A. K. (2006). *Are mutual fund performance measures created equal? An analysis of mutual fund performance and ranking*. Unpublished manuscript, University of North Florida.

Fuertes, A.-M., Miffre, J., & Fernandez-Perez, A. (2015). Commodity strategies based on momentum, term structure, and idiosyncratic volatility. *Journal of Futures Markets*, *35*(3), 274–297.

Fuertes, A.-M., Miffre, J., & Rallis, G. (2010). Tactical allocation in commodity futures markets: Combining momentum and term structure signals. *Journal of Banking and Finance*, *34*(10), 2530–2548.

Garner, C. (2012). *A trader's first book on commodities: An introduction to the world's fastest growing market* (2nd ed.). Upper Saddle River: Pearson Education.

Gemmill, G., Hwang, S., & Salmon, M. (2005). *Performance measurement with loss aversion*. Unpublished manuscript, Cass Business School.

Gilli, M., & Këllezi, E. (2006). An application of extreme value theory for measuring financial risk. *Computational Economics*, *27*, 207–228.

Goetzmann, W., Ingersoll, J., Spiegel, M., & Welch, I. (2007). Portfolio performance manipulation and manipulation-proof performance measures. *Review of Financial Studies*, *20*(5), 1503–1546.

Gorton, G., & Rouwenhorst, K. G. (2006). Facts and fantasies about commodity futures. *Financial Analysts Journal*, *62*(2), 47–68.

Hull, J. C. (2006). *Options, futures, and other derivatives* (6th ed.). Upper Saddle River: Prentice Hall.

Hwang, S., & Salmon, M. (2002). An analysis of performance measures using copulae. In J. Knight & S. Satchell (eds.), *Performance measurement in finance: Firms, funds and managers*. 160–197. London: Butterworth-Heinemann.

Israelsen, C. (2005). A refinement of the Sharpe ratio and information ratio. *Journal of Asset Management*, *5*(6), 423–427.

Jarque, C. M., & Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, *55*(2), 163–172.

Jensen, M. C. (1967). Random walks: Reality or myth—Comment. *Financial Analysts Journal*, *23*(6), 77–85.

Jondeau, E., & Rockinger, M. (2003). Conditional volatility, skewness, and kurtosis: Existence, persistence, and comovements. *Journal of Economic Dynamics and Control*, *27*(10), 1699–1737.

Kallir, I., & Sonsino, D. (2009). The neglect of correlation in allocation decisions. *Southern Economic Journal*, *75*(4), 1045–1066.

Krimm, S., Scholz, H., & Wilkens, M. (2012). The Sharpe ratio's market climate bias: Theoretical and empirical evidence from US equity mutual funds. *Journal of Asset Management*, *13*(4), 227–242.

Kristjanpoller, W., & Minutolo, M. C. (2015). Gold price volatility: A forecasting approach using the Artificial Neural Network-GARCH model. *Expert Systems with Applications*, *42*(20), 7245–7251.

León, Á., Rubio, G., & Serna, G. (2005). Autoregressive conditional volatility, skewness and kurtosis. *Quarterly Review of Economics and Finance, 45*(4–5), 599–618.

Meyer, J. (1987). Two-moment decision models and expected utility maximization. *American Economic Review*, *77*(3), 421–430.

Meyer, J., & Rasche, R. H. (1992). Sufficient conditions for expected utility to imply mean-standard deviation rankings: Empirical evidence concerning the location and scale condition. *Economic Journal*, *102*(410), 91–106.

Miffre, J., & Rallis, G. (2007). Momentum strategies in commodity futures markets. *Journal of Banking and Finance*, *31*(6), 1863–1886.

O'Connor, F., Lucey, B., Batten, J., & Baur, D. (2015). The financial economics of gold—A survey. *International Review of Financial Analysis*, *41*, 186–205.

Ohlson, J., & Rosenberg, B. (1982). Systematic risk of the CRSP equal-weighted common stock index: A history estimated by stochastic parameter regression. *Journal of Business*, *55*(1), 121–145.

Ornelas, J. R. H., Silva Júnior, A. F., & Fernandes, J. L. B. (2012). Yes, the choice of performance measure does matter for ranking of US mutual funds. *International Journal of Finance and Economics*, *17*(1), 61–72.

Pan, M.-S., Liu, Y. A., & Roth, H. J. (1999). Common stochastic trends and volatility in Asian-Pacific equity markets. *Global Finance Journal*, *10*(2), 161–172.

Pedersen, C. S., & Rudholm-Alfvin, T. (2003). Selecting a risk-adjusted shareholder performance measure. *Journal of Asset Management*, *4*(3), 152–172.

Pérignon, C., & Smith, D. R. (2010). The level and quality of Value-at-Risk disclosure by commercial banks. *Journal of Banking and Finance*, *34*(2), 362–377.

Pindyck, R. S. (2001). The dynamics of commodity spot and futures markets: A primer.

Energy Journal, *22*(3), 1–29.

Plantinga, A., & De Groot, S. (2001). Risk-adjusted performance measures and implied risk-attitudes. *Journal of Performance Measurement*, *6*(2), 9–19.

Schuhmacher, F., & Auer, B. R. (2014). Sufficient conditions under which SSD- and MR-efficient sets are identical. *European Journal of Operational Research*, *239*(3), 756–763.

Schuhmacher, F., & Eling, M. (2011). Sufficient conditions for expected utility to imply drawdown-based performance rankings. *Journal of Banking and Finance*, *35*(9), 2311–2318.

Schuhmacher, F., & Eling, M. (2012). A decision-theoretic foundation for reward-to-risk performance measures. *Journal of Banking and Finance*, *36*(7), 2077–2082.

Schuster, M., & Auer, B. R. (2012). A note on empirical Sharpe ratio dynamics. *Economics Letters*, *116*(1), 124–128.

Sévi, B. (2015). Explaining the convenience yield in the WTI crude oil market using realized volatility and jumps. *Economic Modelling*, *44*, 243–251.

Shachmurove, Y. (2011). A historical overview of financial crises in the United States. *Global Finance Journal*, *22*(3), 217–231.

Shukla, R., & Singh, S. (1997). A performance evaluation of global equity mutual funds: Evidence from 1988–1995. *Global Finance Journal*, *8*(2), 279–293.

Sinn, H. W. (1983). *Economic decisions under uncertainty*. Heidelberg: Physica-Verlag.

Szakmary, A. C., Shen, Q., & Sharma, S. C. (2010). Trend-following trading strategies in commodity futures: A re-examination. *Journal of Banking and Finance*, *34*(2), 409–426.

Tang, K., & Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, *68*(5), 54–74.

Xing, K. (2017). Macroeconomic conditions, corporate defaults, and economic recessions. Doctoral dissertation, University of Nottingham.

Zakamouline, V. (2011). The performance measure you choose influences the evaluation of hedge funds. *Journal of Performance Measurment 15*(3), 48–64.

Zakamouline, V. (2014). Portfolio performance evaluation with loss aversion. *Quantitative Finance*, *14*(4), 699–710.

Table 1

Descriptive statistics (full sample, futures and spot)

	Futures							Spot						
	Min	Max	Mean	SD	Skew	Kurt	JB	Min	Max	Mean	SD	Skew	Kurt	JB
Energy														
Crude oil (Brent)	-10.479	12.881	0.002	2.080	-0.112	6.021	13.690	-10.479	12.879	0.013	2.075	-0.069	6.123	14.580
Crude oil (WTI)	-13.065	13.341	-0.025	2.251	-0.141	5.753	11.430	-13.065	13.341	0.011	2.243	-0.108	5.872	12.380
Gas oil	-9.651	10.732	0.009	1.899	0.044	5.105	6.628	-9.657	10.732	0.015	1.889	0.062	5.113	6.684
Heating oil	-9.680	10.067	-0.002	2.084	-0.001	4.820	4.945	-9.678	10.068	0.015	2.082	-0.012	4.848	5.097
Natural gas	-14.641	18.768	-0.137	2.989	0.205	4.878	5.512	-14.645	18.760	-0.009	3.033	0.247	5.002	6.346
Unleaded gasoline	-11.369	12.972	0.017	2.305	-0.200	5.176	7.304	-11.181	12.971	0.018	2.320	-0.154	5.206	7.406
Precious metals														
Gold	-9.810	8.584	0.034	1.196	-0.391	7.818	35.560	-9.811	8.583	0.037	1.198	-0.391	7.799	35.280
Silver	-19.489	12.469	0.025	2.122	-0.902	9.705	71.940	-19.489	12.470	0.028	2.121	-0.901	9.700	71.850
Industrial metals														
Aluminum	-8.253	5.927	-0.018	1.409	-0.283	5.235	7.934	-8.272	5.926	-0.003	1.409	-0.283	5.236	7.944
Copper	-10.397	11.900	0.040	1.801	-0.147	7.004	24.060	-10.382	11.898	0.027	1.802	-0.146	7.017	24.210
Lead	-13.112	12.835	0.031	2.157	-0.216	6.209	15.650	-13.033	12.832	0.028	2.157	-0.216	6.200	15.560
Nickel	-18.256	13.158	0.011	2.402	-0.159	6.310	16.500	-18.224	13.154	0.004	2.402	-0.157	6.302	16.420
Zinc	-11.133	9.853	0.004	1.989	-0.149	5.563	9.935	-11.133	9.926	0.016	1.989	-0.158	5.576	10.050
Agriculture														
Cocoa	-10.014	9.098	0.010	1.843	-0.327	5.884	13.050	-10.006	8.984	0.016	1.844	-0.335	5.880	13.050
Coffee	-11.255	12.080	-0.025	2.045	0.156	5.008	6.160	-11.258	12.050	0.022	2.047	0.138	4.966	5.886
Corn	-8.128	8.670	-0.028	1.801	0.067	5.008	6.046	-8.124	8.663	0.009	1.817	0.047	4.925	5.546
Cotton	-7.123	6.939	-0.022	1.738	-0.083	4.231	2.301	-7.130	6.940	0.007	1.763	-0.081	4.192	2.161
Soybeans	-7.341	6.431	0.035	1.563	-0.220	5.058	6.608	-7.342	6.427	0.016	1.592	-0.246	4.960	6.096
Sugar	-12.369	8.553	-0.006	2.053	-0.254	5.058	6.708	-12.368	8.556	0.014	2.074	-0.248	5.048	6.623
Wheat (Chicago)	-9.973	8.793	-0.041	2.012	0.078	4.726	4.485	-9.972	8.790	0.007	2.016	0.062	4.710	4.387
Wheat (Kansas)	-8.991	8.097	-0.017	1.831	0.065	4.617	3.927	-8.994	8.096	0.009	1.830	0.067	4.625	3.968
Livestock														
Feeder cattle	-5.997	4.255	0.004	0.923	-0.235	4.406	3.282	-6.004	4.251	0.012	0.934	-0.250	4.427	3.412
Lean hogs	-6.409	5.721	-0.054	1.490	-0.048	3.860	1.118	-6.620	7.331	0.006	1.624	0.058	3.958	1.388
Live cattle	-6.359	3.694	-0.006	0.938	-0.185	4.724	4.641	-6.363	3.682	0.011	0.964	-0.200	4.554	3.843

For the period from January 7, 2002, to March 31, 2016, this table reports the minimum, maximum, mean, standard deviation, skewness, kurtosis, and Jarque-Bera (JB) test statistic (which, for better visualization, has been divided by 100) for the daily excess returns of futures and spot commodity subindices of the S&P GSCI. The returns are given in percent. All JB test statistics are significant at the 1% level.

Table 2

Performance measures

No.	Performance measure	Reward measure	Risk measure
Classic			
(1)	Sharpe ratio	μ	σ
Based on drawdowns			
(2)	Calmar ratio	μ	MDD
(3)	Sterling ratio	μ	$K^{-1}\Sigma_{k=1}^{K} CDD_k$
(4)	Burke ratio	μ	$[\Sigma_{k=1}^{K} CDD_k^2]^{1/2}$
(5)	Pain ratio	μ	$T^{-1}\Sigma_{t=1}^{T} DDP_t$
(6)	Martin ratio	μ	$[T^{-1}\Sigma_{t=1}^{T} DDP_t^2]^{1/2}$
Based on partial moments			
(7)	Omega ratio	μ	LPM_1
(8)	Sortino ratio	μ	$LPM_2^{1/2}$
(9)	Kappa 3 ratio	μ	$LPM_3^{1/3}$
(10)	Upside potential ratio	HPM_1	$LPM_2^{1/2}$
Based on the Value-at-Risk			
(11)	Excess return on Value-at-Risk	μ	VaR_α
(12)	Conditional Sharpe ratio	μ	$CVaR_\alpha$
(13)	Modified Sharpe ratio	μ	$MVaR_\alpha$

This table (reproduced from Auer, 2015a) summarizes the reward-to-risk ratios applied in our study. $\mu = T^{-1}(R_1 + \cdots + R_T)$ and $\sigma = [T^{-1}\{(R_1 - \mu)^2 + \cdots + (R_T - \mu)^2\}]^{1/2}$ are the mean and the standard deviation of the excess returns $R_t = r_t - r_{ft}$, $t = 1, \ldots, T$ of a given commodity, where r_t is the daily log return and r_{ft} is the corresponding risk-free rate. MDDdenotes the maximum drawdown (the largest negative cumulative excess return), CDD_k a continuous drawdown (the k-th largest negative cumulative excess return that is not interrupted by a positive excess return) and DDP_t the drawdown from a previous peak (a negative cumulative excess return from the previous peak). K is the number of continuous drawdowns incorporated in the calculation. The signs of the drawdowns are dropped to generate positive risk measures. $HPM_m = T^{-1}\sum_{t=1}^{T} max(R_t, 0)^m$ and $LPM_m = T^{-1}\sum_{t=1}^{T} max(-R_t, 0)^m$ are higher and lower partial moments of order m. VaR_α is the (historical simulation) Value-at-Risk, i.e., the α-quantile of the excess return distribution. The (historical simulation) conditional VaR is estimated as $CVaR_\alpha = B^{-1}\Sigma_{R_t \le -VaR_\alpha} - R_t$, where B is the number of excess returns fulfilling the summation condition. The modified VaR is estimated as $MVaR_\alpha = -[\mu + \sigma\{z_\alpha + (z_\alpha^2 - 1)\gamma/6 + (z_\alpha^3 - 3z_\alpha)\kappa/24 - (2z_\alpha^3 - 5z_\alpha)\gamma^2/36\}]$, where $z_\alpha$is the α-quantile of the standard normal distribution and γ and κ denote skewness and excess kurtosis of the excess return distribution, respectively.

Table 3
Commodity rankings (full sample, futures and spot)

	Futures													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy														
Crude oil (Brent)	12	12	12	12	12	12	12	12	12	18	12	12	12	12.5
Crude oil (WTI)	21	21	21	21	21	21	21	21	21	21	21	21	21	21.0
Gas oil	9	10	10	10	9	9	9	8	8	4	9	8	8	8.5
Heating oil	13	14	13	13	13	13	13	13	13	3	13	13	13	12.3
Natural gas	24	24	24	24	24	24	24	24	24	23	24	24	24	23.9
Unleaded gasoline	6	7	6	6	6	6	6	6	6	7	6	6	6	6.2
Precious metals														
Gold	1	1	1	1	1	1	1	1	1	10	1	1	1	1.7
Silver	5	5	5	5	4	5	5	5	5	24	5	5	5	6.4
Industrial metals														
Aluminum	16	16	16	16	16	16	16	16	16	20	16	16	16	16.3
Copper	3	3	3	3	2	3	2	3	3	13	2	3	2	3.5
Lead	4	4	4	4	5	4	4	4	4	14	4	4	4	4.8
Nickel	8	9	7	7	10	10	8	9	9	9	8	9	9	8.6
Zinc	11	11	11	11	11	11	11	11	11	16	11	11	11	11.4
Agriculture														
Cocoa	7	6	8	8	7	7	7	7	7	19	7	7	7	8.0
Coffee	20	19	19	19	18	18	20	20	19	8	20	19	19	18.3
Corn	19	18	20	20	19	19	19	19	20	15	19	20	20	19.0
Cotton	18	20	18	18	20	20	18	18	18	17	18	18	18	18.4
Soybeans	2	2	2	2	3	2	3	2	2	1	3	2	3	2.2
Sugar	15	15	15	15	15	15	15	15	15	11	15	15	15	14.7
Wheat (Chicago)	23	22	22	22	22	22	22	22	23	6	22	22	22	20.9
Wheat (Kansas)	17	17	17	17	17	17	17	17	17	2	17	17	17	15.8
Livestock														
Feeder cattle	10	8	9	9	8	8	10	10	10	5	10	10	10	9.0
Lean hogs	22	23	23	23	23	23	23	23	22	22	23	23	23	22.8
Live cattle	14	13	14	14	14	14	14	14	14	12	14	14	14	13.8

	Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy														
Crude oil (Brent)	15	17	13	13	13	14	15	15	15	20	15	15	15	15.0
Crude oil (WTI)	17	18	17	18	16	16	17	17	18	16	18	18	17	17.2
Gas oil	11	15	15	15	12	12	11	11	9	10	11	9	9	11.5
Heating oil	13	13	12	14	11	11	13	13	13	7	13	12	13	12.2
Natural gas	24	24	24	24	24	24	24	24	24	5	24	24	24	22.5
Unleaded gasoline	12	10	14	12	8	9	12	12	12	13	12	13	12	11.6
Precious metals														
Gold	1	1	1	1	1	1	1	1	1	14	1	1	1	2.0
Silver	3	6	5	5	5	5	3	4	5	24	3	5	4	5.9
Industrial metals														
Aluminum	23	23	23	23	23	23	23	23	23	18	23	23	23	22.6
Copper	2	5	3	3	2	2	2	2	2	22	2	2	2	3.9
Lead	4	9	4	4	7	6	4	3	4	21	4	4	3	5.9
Nickel	22	22	22	22	22	22	22	22	22	17	22	22	22	21.6
Zinc	10	12	8	7	15	15	10	10	10	19	10	10	10	11.2
Agriculture														
Cocoa	9	3	10	9	9	8	9	9	11	23	9	11	11	10.1
Coffee	7	8	2	2	6	7	7	7	6	3	6	7	6	5.7
Corn	16	14	19	19	17	17	16	16	16	8	17	16	16	15.9
Cotton	19	21	18	17	21	21	19	19	19	12	19	19	19	18.7
Soybeans	8	7	9	8	10	10	8	8	8	15	8	8	8	8.8
Sugar	14	11	11	11	14	13	14	14	14	11	14	14	14	13.0
Wheat (Chicago)	20	19	20	20	20	20	20	20	20	2	20	20	20	18.5
Wheat (Kansas)	18	16	16	16	19	19	18	18	17	1	16	17	18	16.1
Livestock														
Feeder cattle	5	4	6	6	3	3	5	5	3	9	5	3	5	4.8
Lean hogs	21	20	21	21	18	18	21	21	21	4	21	21	21	19.2
Live cattle	6	2	7	10	4	4	6	6	7	6	7	6	7	6.0

For the period from January 7, 2002, to March 31, 2016, and our different performance measures (as numbered in Table 2), this table reports the rankings of our 24 futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

Table 4

Rank correlations and ranking differences (full sample, futures)

	Rank correlations		Ranking differences			
	τ	ρ	Min	Max	MAD	SDAD
(2)	0.9493	0.9922	-2	2	0.5833	0.6539
(3)	0.9710	0.9965	-1	1	0.3333	0.4815
(4)	0.9710	0.9965	-1	1	0.3333	0.4815
(5)	0.9348	0.9904	-2	2	0.5833	0.7755
(6)	0.9493	0.9922	-2	2	0.4167	0.7755
(7)	0.9855	0.9983	-1	1	0.1667	0.3807
(8)	0.9855	0.9983	-1	1	0.1667	0.3807
(9)	0.9855	0.9983	-1	1	0.1667	0.3807
(10)	0.1884	0.2552	-17	19	6.4167	5.6099
(11)	0.9855	0.9983	-1	1	0.1667	0.3807
(12)	0.9783	0.9974	-1	1	0.2500	0.4423
(13)	0.9710	0.9965	-1	1	0.3333	0.4815

Using the futures-based rankings of Table 3, this table presents the Kendall (τ) and Spearman (ρ) rank correlations between the Sharpe ratio and our 12 alternative performance measures. All rank correlation coefficients are significant at the 1% level except the upside potential ratio, which is insignificant. The table also provides descriptive statistics of the ranking differences between our 12 alternative performance measure and the Sharpe ratio: the minima (Min) and maxima (Max) of the differences as well as the mean absolute difference (MAD) and the standard deviation of absolute differences (SDAD).

Table 5

Ranking differences for top five investments (full sample, futures)

	Rank correlations		Ranking differences			
	τ	ρ	Min	Max	MAD	SDAD
(2)	1.0	1.0	0	0	0.00	0.00
(3)	1.0	1.0	0	0	0.00	0.00
(4)	1.0	1.0	0	0	0.00	0.00
(5)	0.6	0.8	-1	1	0.80	0.45
(6)	1.0	1.0	0	0	0.00	0.00
(7)	0.8	0.9	-1	1	0.40	0.55
(8)	1.0	1.0	0	0	0.00	0.00
(9)	1.0	1.0	0	0	0.00	0.00
(10)	0.8	0.9	-1	1	0.40	0.55
(11)	0.8	0.9	-1	1	0.40	0.55
(12)	1.0	1.0	0	0	0.00	0.00
(13)	0.8	0.9	-1	1	0.40	0.55

Like Table 4, this table reports ranking difference statistics, but it concentrates on the five commodities with the highest Sharpe ratios. That is, we identify the commodities with the highest Sharpe ratios, use the alternative performance measures and the Sharpe ratio to rank these assets from 1 to 5, and then calculate the ranking differences.

Table A.1

Commodity rankings (Auer subsample)

	Futures														Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy																												
Crude oil (Brent)	5	8	4	4	8	8	5	5	5	6	5	5	5	5.6	4	9	2	2	4	4	4	4	4	12	4	5	4	4.8
Crude oil (WTI)	11	13	11	11	12	13	11	11	11	14	11	11	11	11.6	6	10	6	6	7	7	7	6	6	14	6	6	6	7.2
Gas oil	4	5	6	6	7	7	4	4	4	1	4	4	4	4.6	2	5	5	4	5	6	2	2	2	3	3	2	2	3.3
Heating oil	9	9	10	10	9	9	9	9	9	2	9	9	9	8.6	5	7	4	5	6	5	5	5	3	1	5	3	5	4.5
Natural gas	24	24	24	24	24	24	24	24	24	22	24	24	24	23.8	23	23	23	23	23	23	23	23	23	5	23	23	23	21.6
Unleaded gasoline	7	6	7	7	5	5	8	7	6	5	7	6	7	6.4	10	11	11	11	8	8	11	10	9	13	10	10	9	10.1
Precious metals																												
Gold	1	1	2	2	1	1	1	1	1	13	1	1	1	2.1	1	1	1	1	1	1	1	1	1	18	1	1	1	2.3
Silver	8	4	8	8	4	4	7	8	8	24	8	8	8	8.2	7	2	8	8	3	3	6	7	10	24	7	7	7	7.6
Industrial metals																												
Aluminum	18	18	18	18	18	18	18	18	18	21	18	18	18	18.2	24	24	24	24	24	24	24	24	24	22	24	24	24	23.8
Copper	2	3	3	3	2	2	2	2	2	8	2	2	2	2.7	3	6	3	3	2	2	3	3	5	19	2	4	3	4.5
Lead	6	7	5	5	6	6	6	6	7	12	6	7	6	6.5	11	12	7	7	11	11	9	11	11	20	11	11	10	10.9
Nickel	10	10	9	9	10	10	10	10	10	7	10	10	10	9.6	20	21	19	19	22	22	20	20	20	17	20	20	20	20.0
Zinc	14	14	14	14	14	14	14	14	14	19	14	14	14	14.4	17	20	18	18	20	21	17	18	19	21	18	19	18	18.8
Agriculture																												
Cocoa	12	11	13	13	11	11	12	12	12	20	12	12	12	12.5	21	15	21	21	19	19	21	21	21	23	21	21	21	20.4
Coffee	21	21	21	21	20	20	21	21	21	17	21	21	21	20.5	15	16	12	12	13	13	15	15	15	7	14	14	14	13.5
Corn	20	19	20	20	19	19	20	20	20	15	20	20	20	19.4	19	17	20	20	16	16	18	19	18	9	19	18	19	17.5
Cotton	19	20	19	19	21	21	19	19	19	16	19	19	19	19.2	14	18	14	13	18	18	14	14	14	11	15	15	15	14.8
Soybeans	3	2	1	1	3	3	3	3	3	3	3	3	3	2.6	8	4	9	9	12	12	8	8	7	15	8	9	8	9.0
Sugar	13	12	12	12	13	12	13	13	13	11	13	13	13	12.5	16	13	17	16	14	14	16	16	17	16	16	16	16	15.6
Wheat (Chicago)	22	22	22	22	22	22	22	22	22	10	22	22	22	21.1	18	19	15	15	17	17	19	17	16	4	17	17	17	16.0
Wheat (Kansas)	16	16	16	16	16	16	16	16	16	4	16	16	16	15.1	13	14	10	10	15	15	13	13	13	2	13	13	13	12.1
Livestock																												
Feeder cattle	15	15	15	15	15	15	15	15	15	9	15	15	15	14.5	9	8	13	14	9	9	10	9	8	8	9	8	11	9.6
Lean hogs	23	23	23	23	23	23	23	23	23	23	23	23	23	23.0	22	22	22	22	21	20	22	22	22	6	22	22	22	20.5
Live cattle	17	17	17	17	17	17	17	17	17	18	17	17	17	17.1	12	3	16	17	10	10	12	12	12	10	12	12	12	11.5

For the period from January 7, 2002, to September 30, 2013, and our different performance measures (as numbered in Table 2), this table reports the rankings of our

24 futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

Table A.2

Commodity rankings (S1—Argentina crisis subsample)

	Futures														Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy																												
Crude oil (Brent)	6	6	6	6	7	6	6	5	5	11	6	6	5	6.2	15	12	12	12	13	12	15	15	14	16	16	14	14	13.8
Crude oil (WTI)	7	5	5	4	5	4	7	6	6	10	7	5	6	5.9	10	11	10	10	9	9	10	10	10	13	14	10	10	10.5
Gas oil	11	9	8	8	9	9	11	11	10	8	11	10	10	9.6	9	9	9	8	6	7	9	9	8	7	9	8	9	8.2
Heating oil	10	7	9	9	8	7	8	10	11	14	10	11	11	9.6	7	7	5	5	4	4	7	8	9	14	8	9	8	7.3
Natural gas	8	8	7	7	10	10	9	8	8	5	8	7	8	7.9	3	3	1	1	3	3	3	3	3	3	3	3	3	2.7
Unleaded gasoline	13	12	13	13	12	12	13	13	13	15	13	13	13	12.9	17	15	15	15	15	15	17	17	17	15	17	17	17	16.1
Precious metals																												
Gold	3	3	3	3	4	3	3	3	3	9	3	3	3	3.5	6	6	8	9	7	6	6	6	7	12	5	7	7	7.1
Silver	21	20	21	21	20	20	21	21	21	24	21	21	21	21.0	21	21	21	21	21	21	21	21	22	24	21	22	21	21.4
Industrial metals																												
Aluminum	19	19	19	19	19	19	19	19	19	18	19	19	19	18.9	19	19	19	19	19	19	19	19	19	18	19	19	19	18.9
Copper	15	16	16	16	16	16	16	15	15	7	15	15	15	14.8	16	16	17	17	16	16	16	16	16	8	15	16	16	15.5
Lead	23	23	23	23	23	23	23	23	23	22	23	23	23	22.9	23	24	23	23	24	24	24	24	24	23	23	24	24	23.6
Nickel	4	4	4	5	6	5	5	4	4	4	4	4	4	4.4	11	10	11	11	12	11	12	12	11	10	11	11	12	11.2
Zinc	22	22	22	22	22	22	22	22	22	21	22	22	22	21.9	20	20	20	20	20	20	20	20	20	21	20	20	20	20.1
Agriculture																												
Cocoa	5	11	11	11	3	8	4	7	9	19	5	9	7	8.4	13	14	14	14	10	13	13	14	15	22	12	15	15	14.2
Coffee	14	14	14	14	15	15	14	14	14	12	14	14	14	14	4	4	7	7	8	8	4	4	4	5	4	4	4	5.2
Corn	18	18	18	18	18	18	18	18	18	16	18	18	18	17.8	14	13	13	13	14	14	14	13	13	9	13	13	13	13.0
Cotton	16	15	15	15	14	14	15	16	16	13	16	16	16	15.2	5	5	4	4	5	5	5	5	5	6	6	5	5	5.0
Soybeans	1	1	1	1	1	1	1	1	1	1	1	1	1	1.0	2	1	3	2	2	2	2	2	2	2	2	2	2	2.0
Sugar	9	10	10	10	11	11	10	9	7	3	9	8	9	8.9	24	23	24	24	23	23	23	23	23	17	24	23	23	22.8
Wheat (Chicago)	12	13	12	12	13	13	12	12	12	6	12	12	12	11.8	8	8	6	6	11	10	8	7	6	4	7	6	6	7.2
Wheat (Kansas)	2	2	2	2	2	2	2	2	2	2	2	2	2	2.0	1	2	2	3	1	1	1	1	1	1	1	1	1	1.3
Livestock																												
Feeder cattle	20	21	20	20	21	21	20	20	20	20	20	20	20	20.2	18	18	18	18	18	18	18	18	18	19	18	18	18	18.1
Lean hogs	24	24	24	24	24	24	24	24	24	23	24	24	24	23.9	22	22	22	22	22	22	22	22	21	20	22	21	22	21.7
Live cattle	17	17	17	17	17	17	17	17	17	17	17	17	17	17.0	12	17	16	16	17	17	11	11	12	11	10	12	11	13.3

For the period from January 7, 2002, to November 30, 2002, and our different performance measures (as numbered in Table 2), this table reports the rankings of our

24 futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

Table A.3

Commodity rankings (S2—Growth subsample)

	Futures														Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy																												
Crude oil (Brent)	3	3	2	2	3	2	4	3	3	1	3	3	3	2.7	2	2	2	2	2	1	3	2	1	1	2	1	2	1.8
Crude oil (WTI)	6	5	6	6	6	6	6	6	5	4	6	5	6	5.6	5	4	4	4	5	5	5	5	4	3	5	4	5	4.5
Gas oil	2	2	3	3	4	3	2	2	2	3	2	2	2	2.5	3	1	5	5	4	4	4	3	2	4	3	2	3	3.3
Heating oil	7	8	9	10	10	9	7	7	7	2	7	7	7	7.5	6	5	7	9	7	6	7	6	6	2	6	6	6	6.1
Natural gas	24	24	24	24	24	24	24	24	24	22	24	24	24	23.8	20	20	20	20	20	20	20	20	20	16	20	20	20	19.7
Unleaded gasoline	8	9	11	11	9	8	9	8	8	5	9	8	8	8.5	8	8	9	8	9	9	9	8	8	5	9	8	8	8.2
Precious metals																												
Gold	5	4	5	5	5	5	5	5	6	13	5	6	5	5.7	4	3	3	3	3	3	2	4	5	15	4	5	4	4.5
Silver	11	6	12	12	7	7	8	11	11	24	10	11	11	10.8	9	7	10	10	8	7	8	9	11	24	8	10	10	10.1
Industrial metals																												
Aluminum	12	10	10	9	11	11	12	12	12	15	12	12	12	11.5	12	10	8	7	10	10	12	12	12	19	12	12	12	11.4
Copper	1	1	1	1	1	1	1	1	1	6	1	1	1	1.4	1	6	1	1	1	2	1	1	3	9	1	3	1	2.4
Lead	4	7	4	4	2	4	3	4	4	7	4	4	4	4.2	7	9	6	6	6	8	6	7	7	11	7	7	7	7.2
Nickel	10	12	8	8	8	10	11	10	10	9	8	9	10	9.5	13	16	14	14	11	15	13	13	13	12	13	13	13	13.3
Zinc	13	14	13	13	13	14	13	13	13	16	13	13	13	13.4	14	18	13	12	15	16	14	16	16	20	16	17	16	15.6
Agriculture																												
Cocoa	16	16	16	16	16	16	15	16	16	21	16	16	16	16.3	19	17	19	19	21	19	19	19	19	23	19	19	19	19.3
Coffee	21	19	21	21	21	19	21	21	21	18	21	21	21	20.5	17	12	15	15	16	13	16	17	17	14	17	16	17	15.5
Corn	18	20	18	18	20	21	18	18	18	12	18	18	18	18.1	10	11	11	11	13	14	10	10	9	6	10	9	9	10.2
Cotton	23	23	23	23	23	23	23	23	23	23	23	23	23	23.0	23	23	21	21	23	23	23	23	23	17	23	23	23	22.2
Soybeans	9	11	7	7	12	12	10	9	9	10	11	10	9	9.7	11	15	12	13	17	17	11	11	10	13	11	11	11	12.5
Sugar	19	21	20	20	19	20	19	20	20	19	20	20	20	19.8	18	19	18	18	18	18	18	18	18	18	18	18	18	18.1
Wheat (Chicago)	20	18	19	19	18	18	20	19	19	11	19	19	19	18.3	15	13	17	17	14	12	17	14	14	7	15	14	15	14.2
Wheat (Kansas)	15	15	15	15	15	15	16	15	15	8	15	15	15	14.5	16	14	16	16	12	11	15	15	15	8	14	15	14	13.9
Livestock																												
Feeder cattle	14	13	14	14	14	13	14	14	14	14	14	14	14	13.8	22	21	23	22	19	21	22	22	22	22	22	22	22	21.7
Lean hogs	22	22	22	22	22	22	22	22	22	20	22	22	22	21.8	24	24	24	24	24	24	24	24	24	10	24	24	24	22.9
Live cattle	17	17	17	17	17	17	17	17	17	17	17	17	17	17.0	21	22	22	23	22	22	21	21	21	21	21	21	21	21.5

For the period from December 1, 2002, to August 1, 2008, and our different performance measures (as numbered in Table 2), this table reports the rankings of our

24 futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

Table A.4

Commodity rankings (S3—Lehman Brothers crisis subsample)

	Futures														Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy																												
Crude oil (Brent)	22	22	21	21	22	22	22	22	22	23	22	22	22	21.9	20	19	19	19	20	20	20	20	21	23	20	22	20	20.2
Crude oil (WTI)	23	23	23	23	23	23	23	23	23	24	23	23	23	23.1	23	23	23	23	22	22	23	23	23	24	23	23	23	22.9
Gas oil	20	21	22	22	21	21	20	20	20	17	20	20	20	20.3	22	22	22	22	23	23	22	22	22	18	21	19	22	21.5
Heating oil	21	20	20	20	20	20	21	21	21	19	21	21	21	20.5	21	21	20	20	21	21	21	21	20	20	22	21	21	20.8
Natural gas	24	24	24	24	24	24	24	24	24	20	24	24	24	23.7	24	24	24	24	24	24	24	24	24	7	24	24	24	22.7
Unleaded gasoline	19	19	19	19	17	17	18	19	19	21	18	19	19	18.7	19	20	21	21	19	19	19	19	19	22	19	20	19	19.7
Precious metals																												
Gold	1	1	1	1	1	1	1	1	1	1	1	1	1	1.0	1	1	3	2	1	1	1	1	1	2	1	1	1	1.3
Silver	2	2	2	2	2	2	2	2	2	6	2	2	2	2.3	3	2	4	4	2	2	2	4	4	8	3	4	2	3.4
Industrial metals																												
Aluminum	15	15	15	15	15	15	15	15	15	15	15	15	15	15.0	18	18	18	18	18	18	18	18	18	17	18	18	18	17.9
Copper	7	9	7	7	7	8	7	7	7	9	7	7	7	7.4	7	11	9	9	10	11	7	7	7	14	8	7	7	8.8
Lead	10	10	10	10	10	10	10	10	10	8	10	10	10	9.8	8	7	8	8	7	7	8	8	8	9	7	8	8	7.8
Nickel	8	8	8	8	8	7	8	8	8	7	8	8	8	7.8	11	8	7	7	9	9	9	10	10	12	9	10	9	9.2
Zinc	9	7	9	9	9	9	9	9	9	5	9	9	9	8.5	6	6	6	6	6	6	6	6	6	6	6	6	6	6.0
Agriculture																												
Cocoa	11	11	11	11	11	11	11	11	11	13	11	11	11	11.2	12	12	12	12	12	12	12	12	12	19	12	12	12	12.5
Coffee	5	5	5	5	5	5	5	5	5	4	5	5	5	4.9	5	5	5	5	5	5	5	5	5	4	5	5	5	4.9
Corn	17	16	18	18	16	16	16	16	17	16	16	17	16	16.5	17	17	17	17	17	17	17	17	17	16	17	17	17	16.9
Cotton	3	3	3	3	4	4	3	3	3	3	4	3	3	3.2	4	3	1	1	4	3	3	3	2	3	4	3	3	2.8
Soybeans	6	6	6	6	6	6	6	6	6	14	6	6	6	6.6	16	16	16	16	16	15	15	16	16	21	15	16	16	16.2
Sugar	4	4	4	4	3	3	4	4	4	2	3	4	4	3.6	2	4	2	3	3	4	4	2	3	1	2	2	4	2.8
Wheat (Chicago)	18	17	16	16	19	19	19	18	18	11	19	18	18	17.4	15	15	15	15	15	16	16	15	15	11	16	15	15	14.9
Wheat (Kansas)	14	14	14	14	14	14	14	14	14	10	14	14	14	13.7	14	14	14	14	14	14	14	14	14	10	14	14	14	13.7
Livestock																												
Feeder cattle	12	12	12	12	12	12	12	12	12	12	12	12	12	12.0	10	9	11	11	11	10	10	11	11	13	11	11	11	10.8
Lean hogs	16	18	17	17	18	18	17	17	16	18	17	16	17	17.1	9	10	10	10	8	8	11	9	9	5	10	9	10	9.1
Live cattle	13	13	13	13	13	13	13	13	13	22	13	13	13	13.7	13	13	13	13	13	13	13	13	13	15	13	13	13	13.2

For the period from September 1, 2008, to December 7, 2010, and our different performance measures (as numbered in Table 2), this table reports the rankings of our 24 futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

Table A.5

Commodity rankings (S4—EU crisis subsample)

	Futures														Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy																												
Crude oil (Brent)	2	3	3	3	1	2	2	1	2	1	1	1	1	1.8	5	4	4	4	2	3	3	3	4	3	1	2	3	3.2
Crude oil (WTI)	14	12	13	13	12	12	13	13	13	11	11	13	13	12.5	11	8	9	12	9	8	11	10	10	9	10	10	9	9.7
Gas oil	1	1	1	1	2	1	1	2	1	4	4	4	2	1.9	2	1	1	1	1	1	2	2	2	5	6	5	2	2.4
Heating oil	3	2	2	2	3	3	3	3	3	2	3	3	3	2.7	6	2	3	3	3	2	6	4	3	2	5	4	4	3.6
Natural gas	24	24	24	24	24	24	24	24	24	20	24	24	24	23.7	24	24	23	23	24	24	24	24	23	20	24	23	24	23.4
Unleaded gasoline	5	4	4	4	4	4	4	6	7	7	5	7	5	5.1	3	5	5	5	4	4	4	6	6	7	4	7	6	5.1
Precious metals																												
Gold	20	20	20	20	20	20	20	20	20	21	20	20	20	20.1	20	20	20	20	20	20	20	20	20	22	20	20	20	20.2
Silver	8	9	8	8	10	10	7	8	9	14	13	12	8	9.5	10	11	10	9	12	12	9	11	13	13	13	13	11	11.3
Industrial metals																												
Aluminum	11	8	10	9	9	9	12	11	10	9	10	9	11	9.8	13	10	11	10	10	9	13	13	11	12	12	11	13	11.4
Copper	17	16	16	16	17	17	17	17	16	17	18	16	17	16.7	17	17	17	17	18	17	17	17	17	18	18	17	17	17.2
Lead	13	13	14	14	13	13	14	14	14	13	14	14	14	13.6	14	14	14	14	14	14	14	14	14	14	14	14	14	14.0
Nickel	18	18	18	18	18	18	18	18	18	16	17	18	18	17.8	18	19	18	19	17	18	18	18	18	16	17	18	18	17.8
Zinc	19	19	19	19	19	19	19	19	19	18	19	19	19	18.9	19	18	19	18	19	19	19	19	19	17	19	19	19	18.7
Agriculture																												
Cocoa	22	22	22	22	22	22	22	22	22	23	22	22	22	22.1	22	23	24	24	22	22	22	22	22	23	22	22	22	22.5
Coffee	9	11	9	10	11	11	10	9	8	8	8	8	9	9.3	12	13	13	13	13	13	12	12	12	11	11	12	12	12.2
Corn	7	10	6	5	6	7	8	7	6	5	7	6	7	6.7	9	12	7	7	8	10	8	8	8	8	9	8	7	8.4
Cotton	4	5	7	7	8	6	6	4	4	3	6	2	6	5.2	7	7	12	11	11	11	10	7	7	4	7	3	8	8.1
Soybeans	16	17	17	17	16	16	16	16	17	19	16	17	16	16.6	16	16	16	16	16	15	16	16	16	19	16	16	16	16.2
Sugar	21	21	21	21	21	21	21	21	21	24	21	21	21	21.2	23	22	22	22	23	23	23	23	24	24	23	24	23	23.0
Wheat (Chicago)	23	23	23	23	23	23	23	23	23	22	23	23	23	22.9	21	21	21	21	21	21	21	21	21	21	21	21	21	21.0
Wheat (Kansas)	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	16	15	15	15	15	15	15	15	15.1
Livestock																												
Feeder cattle	6	6	5	6	5	5	5	5	5	6	2	5	4	5.0	4	6	6	6	6	6	5	5	5	6	3	6	5	5.3
Lean hogs	12	14	11	11	14	14	11	12	11	10	12	10	12	11.8	1	3	2	2	5	5	1	1	1	1	2	1	1	2.0
Live cattle	10	7	12	12	7	8	9	10	12	12	9	11	10	9.9	8	9	8	8	7	7	7	9	9	10	8	9	10	8.4

For the period from December 8, 2010, to April 4, 2011, and our different performance measures (as numbered in Table 2), this table reports the rankings of our 24

futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

Table A.6

Commodity rankings (S5—Greek crisis subsample)

	Futures														Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy																												
Crude oil (Brent)	3	3	3	3	3	3	3	3	3	12	3	3	3	3.7	5	5	5	5	5	5	5	5	5	16	5	5	5	5.8
Crude oil (WTI)	10	11	10	11	12	11	10	12	12	16	11	12	12	11.5	9	9	9	9	9	9	9	9	9	15	9	9	9	9.5
Gas oil	4	5	5	5	4	4	4	5	5	3	4	5	5	4.5	6	6	6	6	6	6	6	6	6	5	6	6	6	5.9
Heating oil	8	7	8	8	8	7	8	8	8	13	8	8	8	8.2	7	7	7	7	7	7	7	7	7	12	7	7	7	7.4
Natural gas	24	24	24	24	24	24	24	24	24	24	24	24	24	24.0	24	23	24	24	23	23	24	24	24	23	23	24	24	23.6
Unleaded gasoline	2	1	2	2	2	2	2	2	2	5	2	2	2	2.2	3	3	3	3	3	3	3	3	3	14	3	3	3	3.8
Precious metals																												
Gold	1	2	1	1	1	1	1	1	1	6	1	1	1	1.5	1	1	1	1	2	1	1	1	1	6	1	1	1	1.5
Silver	16	16	16	16	16	16	16	16	17	22	16	17	17	16.7	17	17	19	19	17	17	17	17	20	22	17	19	18	18.2
Industrial metals																												
Aluminum	14	15	14	14	15	15	14	14	14	20	14	14	14	14.7	14	16	14	14	15	15	14	14	13	17	13	13	13	14.2
Copper	12	12	12	12	11	12	11	10	11	10	12	11	10	11.2	11	11	11	11	10	10	10	11	11	11	11	11	11	10.8
Lead	18	17	22	22	17	17	18	18	18	15	20	18	18	18.3	21	19	22	22	19	19	21	21	21	18	21	21	21	20.5
Nickel	21	19	21	21	19	19	20	21	21	18	22	21	21	20.3	22	21	21	21	22	22	22	22	22	19	22	22	22	21.5
Zinc	15	14	15	15	14	14	15	15	15	9	15	15	15	14.3	15	15	16	16	14	14	15	15	15	9	15	15	15	14.5
Agriculture																												
Cocoa	17	18	17	17	18	18	17	17	16	8	17	16	16	16.3	18	20	18	18	18	18	18	18	17	8	18	17	17	17.2
Coffee	19	22	20	20	20	20	19	19	19	21	19	19	19	19.7	20	22	20	20	21	21	20	20	19	21	20	20	20	20.3
Corn	13	13	13	13	13	13	13	13	13	19	13	13	13	13.5	13	13	13	12	13	13	13	13	14	20	14	14	14	13.8
Cotton	23	23	23	23	23	23	23	23	23	23	23	23	23	23.0	23	24	23	23	24	24	23	23	23	24	24	23	23	23.4
Soybeans	6	8	6	6	7	8	6	6	6	4	6	6	6	6.2	4	4	4	4	4	4	4	4	4	4	4	4	4	4.0
Sugar	5	4	4	4	5	5	5	4	4	1	5	4	4	4.2	12	12	12	13	12	12	12	12	12	3	12	12	12	11.4
Wheat (Chicago)	22	21	18	18	22	22	22	22	22	17	21	22	22	20.8	16	14	15	15	16	16	16	16	16	10	16	16	16	15.2
Wheat (Kansas)	20	20	19	19	21	21	21	20	20	14	18	20	20	19.5	19	18	17	17	20	20	19	19	18	13	19	18	19	18.2
Livestock																												
Feeder cattle	7	6	7	7	6	6	7	7	7	2	7	7	7	6.4	2	2	2	2	1	2	2	2	2	1	2	2	2	1.8
Lean hogs	11	10	11	10	10	10	12	11	10	7	10	10	11	10.2	10	10	10	10	11	11	11	10	10	7	10	10	10	10.0
Live cattle	9	9	9	9	9	9	9	9	9	11	9	9	9	9.2	8	8	8	8	8	8	8	8	8	2	8	8	8	7.5

For the period from April 5, 2011, to March 31, 2012, and our different performance measures (as numbered in Table 2), this table reports the rankings of our 24

futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

Table A.7

Commodity rankings (S6—Post-crisis subsample)

	Futures														Spot													
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	Mean
Energy																												
Crude oil (Brent)	23	23	23	22	22	23	23	23	23	24	23	23	23	22.9	24	24	23	23	24	24	24	24	24	24	24	24	24	23.8
Crude oil (WTI)	24	24	24	24	24	24	24	24	24	22	24	24	24	23.8	23	22	21	21	22	23	23	23	23	20	23	23	23	22.3
Gas oil	20	22	22	23	20	22	21	20	20	23	21	20	20	21.1	21	23	24	24	21	22	20	20	20	23	20	20	20	21.4
Heating oil	21	21	21	21	18	19	20	21	21	20	20	21	21	20.4	22	21	20	20	18	21	22	22	22	21	22	22	21	21.1
Natural gas	22	20	20	20	19	18	22	22	22	9	22	22	22	20.0	6	6	5	5	5	5	6	6	6	2	6	6	6	5.4
Unleaded gasoline	16	18	19	19	12	12	16	16	17	18	18	17	16	16.5	20	20	22	22	19	19	21	21	21	18	21	21	22	20.5
Precious metals																												
Gold	6	8	7	7	9	9	6	7	8	19	6	7	7	8.2	8	9	8	8	9	9	8	8	9	19	8	9	8	9.2
Silver	18	16	17	17	21	20	15	18	19	21	17	19	18	18.2	19	18	19	19	23	20	18	19	19	22	19	19	19	19.5
Industrial metals																												
Aluminum	11	11	11	11	11	11	11	10	10	6	10	10	10	10.2	9	8	9	9	8	8	9	9	8	4	9	8	9	8.2
Copper	10	10	10	10	10	10	10	11	11	17	11	11	11	10.9	14	15	12	12	14	14	14	14	14	17	15	14	14	14.1
Lead	7	6	6	6	6	6	7	6	6	2	7	6	6	5.9	7	5	7	7	7	6	7	7	7	3	7	7	7	6.5
Nickel	17	17	15	15	13	15	18	17	16	16	16	16	17	16.0	18	19	17	18	17	17	19	18	18	15	18	18	18	17.7
Zinc	5	5	5	5	5	5	5	5	5	4	5	5	5	4.9	4	4	4	4	4	4	4	4	4	5	4	4	4	4.1
Agriculture																												
Cocoa	1	1	1	1	1	1	1	1	1	1	1	1	1	1.0	1	1	1	1	1	1	1	1	1	1	1	1	1	1.0
Coffee	19	15	18	18	17	17	19	19	18	11	19	18	19	17.5	16	10	13	13	10	10	16	16	16	9	16	16	16	13.6
Corn	12	12	13	13	14	13	12	12	12	13	12	12	12	12.5	17	17	18	17	20	18	17	17	17	14	17	17	17	17.2
Cotton	9	7	9	9	8	7	9	9	9	15	9	9	9	9.1	13	11	15	15	11	11	13	13	13	16	13	13	13	13.1
Soybeans	2	2	2	2	2	2	2	2	2	3	2	2	2	2.1	11	14	16	16	15	15	11	12	12	13	10	12	12	13.0
Sugar	15	19	16	16	23	21	17	15	15	14	15	15	15	16.6	15	16	14	14	16	16	15	15	15	12	14	15	15	14.8
Wheat (Chicago)	13	13	12	12	16	14	13	13	13	7	13	13	13	12.7	10	12	10	10	12	12	10	10	10	8	11	10	10	10.4
Wheat (Kansas)	14	14	14	14	15	16	14	14	14	8	14	14	14	13.8	12	13	11	11	13	13	12	11	11	7	12	11	11	11.4
Livestock																												
Feeder cattle	4	4	4	4	4	4	4	4	4	12	4	4	4	4.6	3	3	3	3	3	3	3	3	3	11	3	3	3	3.6
Lean hogs	8	9	8	8	7	8	8	8	7	10	8	8	8	8.1	5	7	6	6	6	7	5	5	5	10	5	5	5	5.9
Live cattle	3	3	3	3	3	3	3	3	3	5	3	3	3	3.2	2	2	2	2	2	2	2	2	2	6	2	2	2	2.3

For the period from April 1, 2012, to March 31, 2016, and our different performance measures (as numbered in Table 2), this table reports the rankings of our 24

futures-based and our 24 spot-based commodity investments, where 1 (24) resembles the best (worst) investment performance. For each investment subgroup, the last column reports the mean ranking across all performance measures.

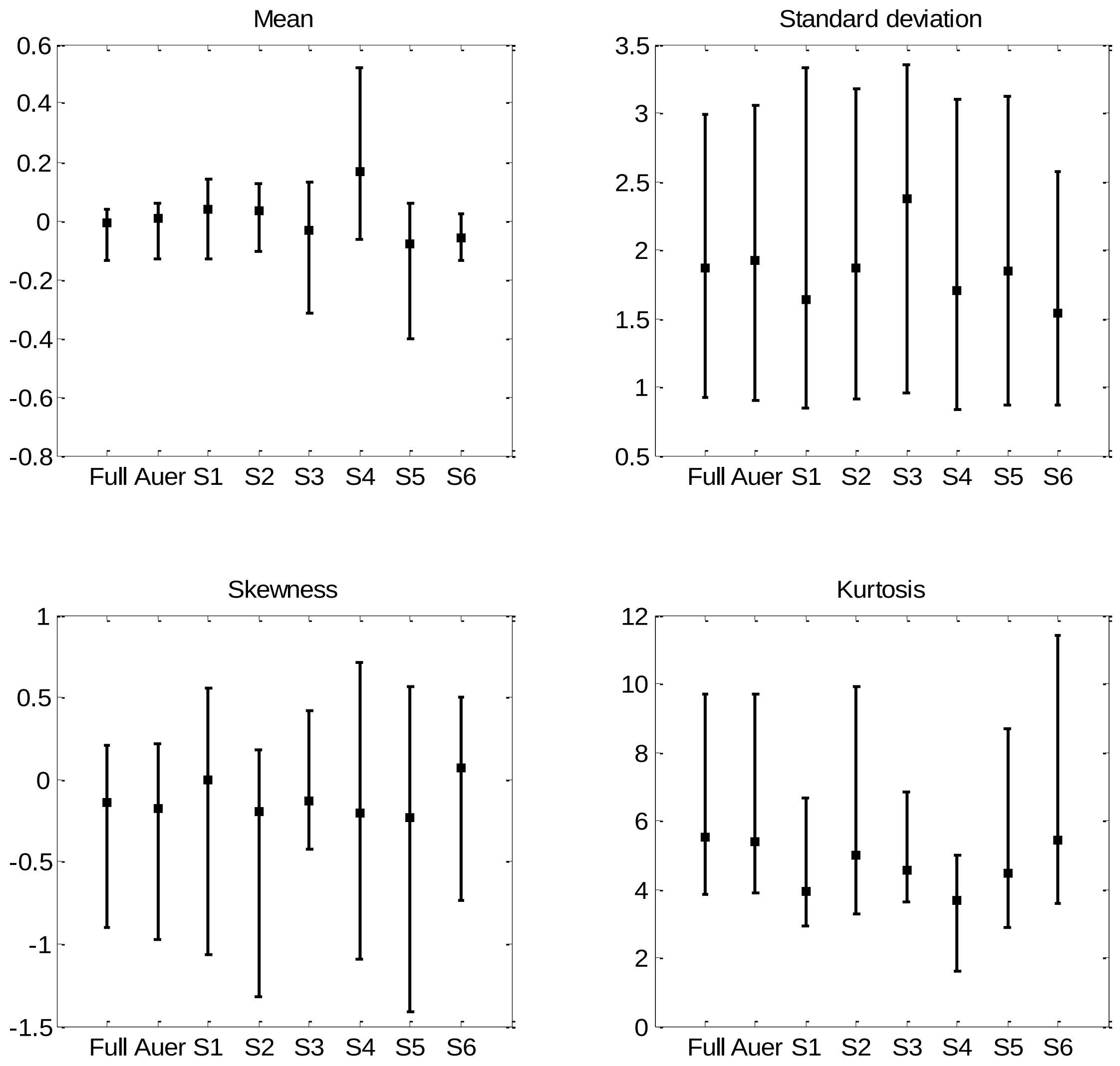

Fig. 1. Descriptive statistics (subsamples, futures).

For our 24 commodity futures indices, this figure illustrates the cross-sectional averages (bold dots) of the sample means, standard deviations, skewness values, and kurtosis values within each of our sample specifications. The highest and lowest realizations of these metrics are represented by a band around each average. Sample abbreviations are used as follows: Full: January 7, 2002–March 31, 2016; Auer: January 7, 2002–September 30, 2013; S1: January 7, 2002–November 30, 2002; S2: December 1, 2002–August 1, 2008; S3: September 1, 2008–December 7, 2010; S4: December 8, 2010–April 4, 2011; S5: April 5, 2011–March 31, 2012; S6: April 1, 2012–March 31, 2016.

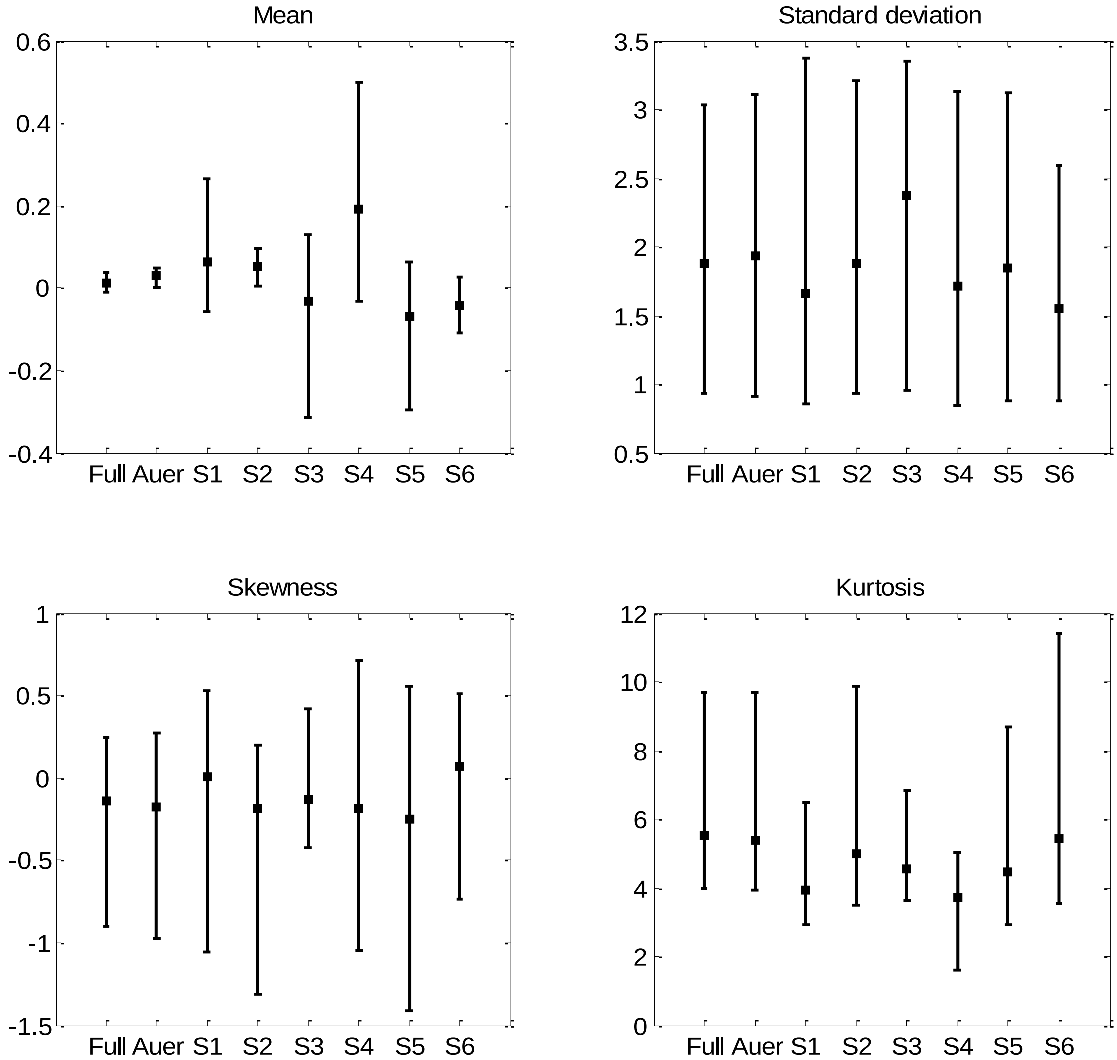

Fig. 2. Descriptive statistics (subsamples, spot).

For our 24 commodity spot indices, this figure illustrates the cross-sectional averages (bold dots) of the sample means, standard deviations, skewness values, and kurtosis values within each of our subsample specifications. The highest and lowest realizations of these metrics are represented by a band around each average. Sample abbreviations are used as follows: Full: January 7, 2002–March 31, 2016; Auer: January 7, 2002–September 30, 2013; S1: January 7, 2002–November 30, 2002; S2: December 1, 2002–August 1, 2008; S3: September 1, 2008–December 7, 2010; S4: December 8, 2010–April 4, 2011; S5: April 5, 2011–March 31, 2012; S6: April 1, 2012–March 31, 2016.

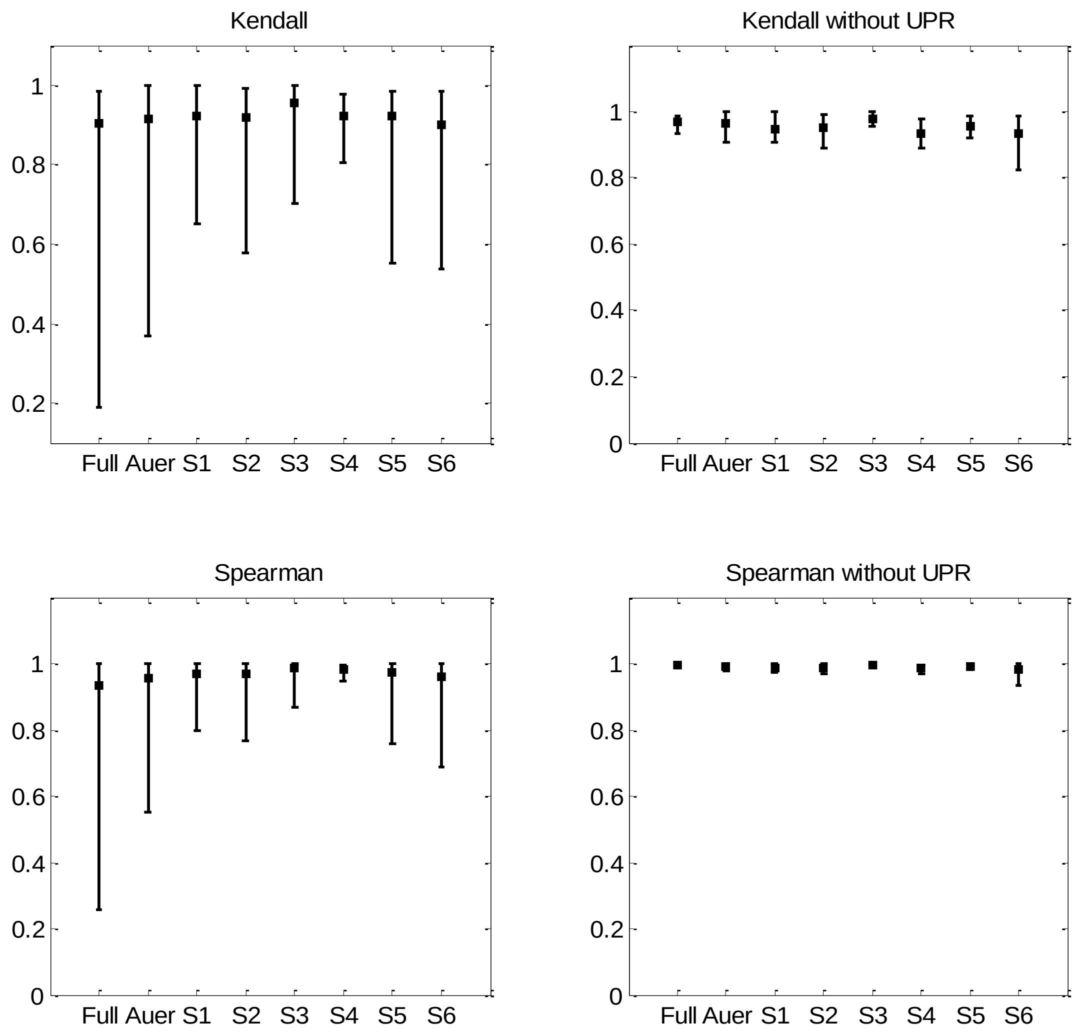

Fig. 3. Rank correlations (subsamples, futures).

Similar to the visualization in Figures 1 and 2, for each of our subsamples, the left side of this figure presents the means, minima, and maxima of Kendal's and Spearman's rank correlation coefficients between our 12 alternative performance measures and the Sharpe ratio. The right side shows the results when the correlation values for the upside potential ratio (UPR) are excluded.

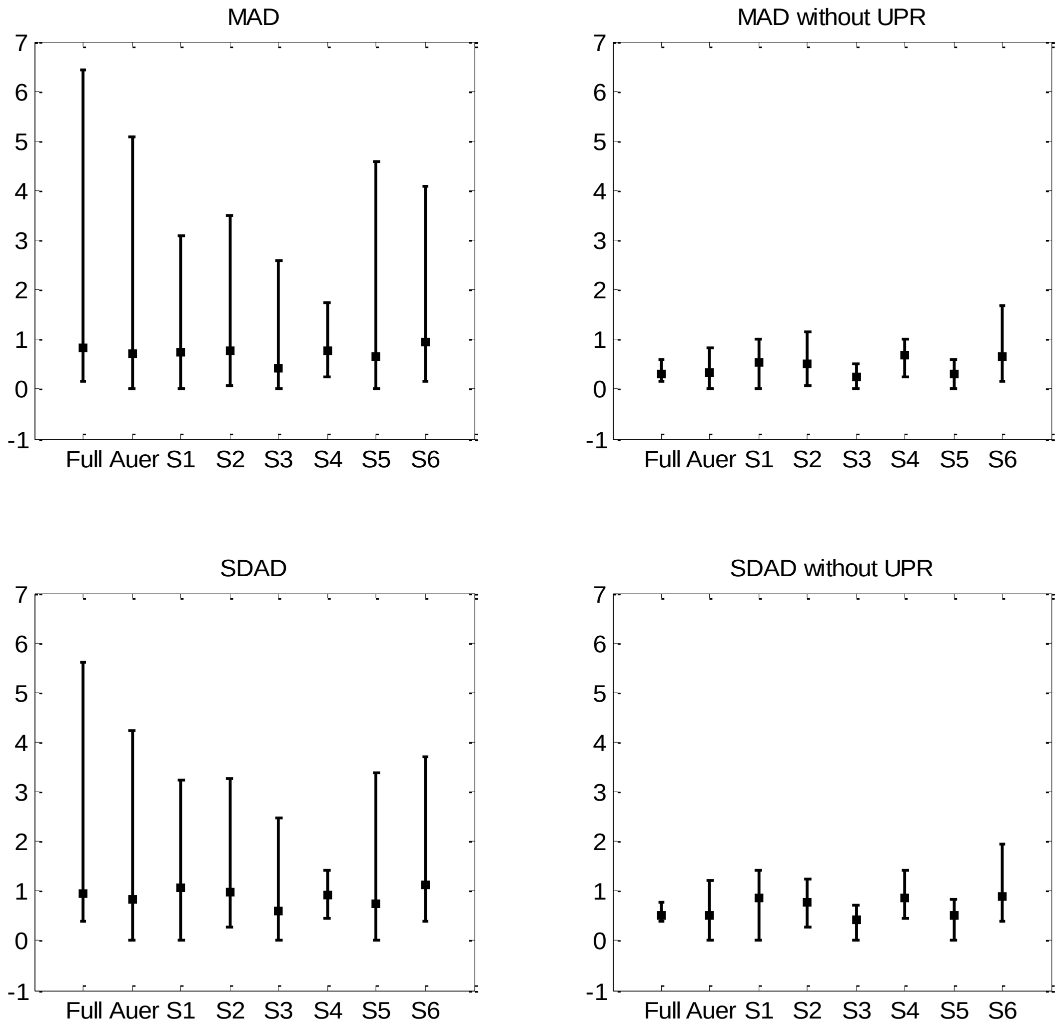

Fig. 4. Ranking differences (subsamples, futures).

Similar to the visualization in Figures 1 and 2, for each of our subsamples, the left side of this figure presents the means, minima, and maxima of mean absolute differences (MAD) and standard deviations of absolute differences (SDAD) between the rankings of our 12 alternative performance measures and the Sharpe ratio. The right side shows the results when the MAD and SDAD values for the upside potential ratio (UPR) are excluded.

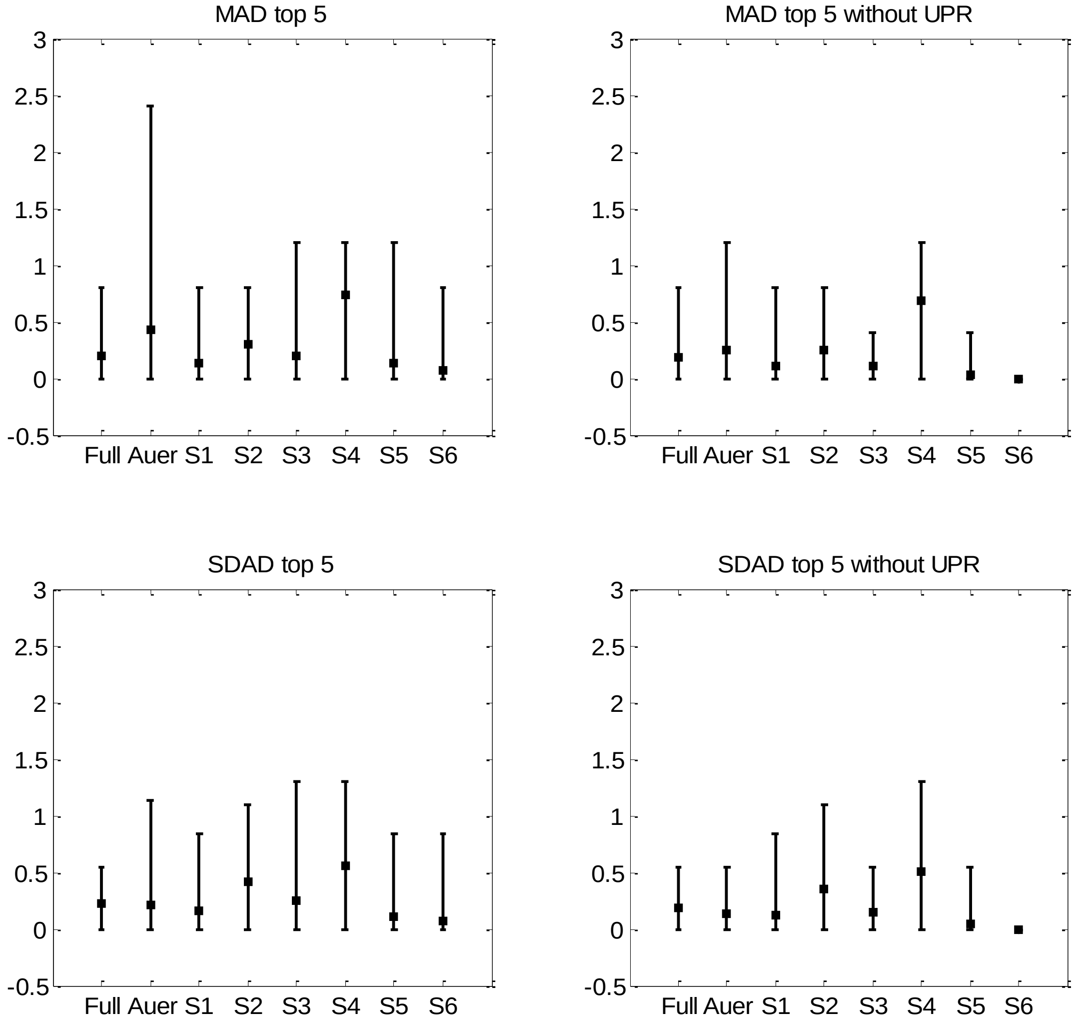

Fig. 5. Ranking differences for top five investments (subsamples, futures).

This figure is similar to Figure 4. While Figure 4 includes all commodities, this figure focuses on the differences for the five commodities with the highest Sharpe ratios.